NBER WORKING PAPER SERIES

BRINGING REAL MARKET PARTICIPANTS' REAL PREFERENCES INTO THE LAB: AN EXPERIMENT THAT CHANGED THE COURSE ALLOCATION MECHANISM AT WHARTON

Eric Budish
Judd B. Kessler

Working Paper 22448
http://www.nber.org/papers/w22448

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2016, Revised December 2018

The authors thank Gérard Cachon, without whom this study would never have been possible, and the Wharton School, in particular the Course Allocation Redesign Team and Wharton Computing. We thank Natalia Drozdoff, Adriaan Ten Kate and Xan Vongsathorn for excellent research assistance. We also thank Mohammad Akbarpour, Eduardo Azevedo, Peter Cramton, Stefano DellaVigna, Clayton Featherstone, Alex Frankel, Emir Kamenica, Scott Kominers, Robin Lee, Stephen Leider, John List, Paul Milgrom, Joshua Mollner, Muriel Niederle, Canice Prendergast, Jesse Shapiro, Alvin Roth and Glen Weyl, as well as seminar participants at Boston University, the Stony Brook Workshop on Experimental Game Theory, ESA Santa Cruz, University of Michigan, Stanford, Wharton, NBER Market Design, Chicago, AMMA 2015, MSR Designing the Digital Economy, Boston College, Princeton and the University of Virginia. Disclosure: the mechanism design in Budish (2011) and the computational procedure in Othman, Budish and Sandholm (2010) are in the public domain. Wharton funded the implementation of the A-CEEI mechanism at Wharton, creating intellectual property owned by Wharton. In 2017, Wharton spun out this intellectual property into a startup, Cognomos, of which Budish was made a director at its founding and has an equity stake. Wharton had no right of prior review of the present study. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Bringing Real Market Participants' Real Preferences into the Lab: An Experiment that Changed the Course Allocation Mechanism at Wharton

Eric Budish and Judd B. Kessler

NBER Working Paper No. 22448
July 2016, Revised December 2018
JEL No. C78,C9,D47

ABSTRACT

This paper reports on an experimental test of a new market design that is attractive in theory but makes the common and potentially unrealistic assumption that "agents report their type"; that is, that market participants can perfectly report their preferences to the mechanism. This concern about preference reporting led to a novel experimental design that brought real market participants' real preferences into the lab, as opposed to endowing experimental subjects with artificial preferences as is typical in market design laboratory experiments. The experiment found that market participants were able to report their preferences "accurately enough" to realize efficiency and fairness benefits of the mechanism, though preference reporting mistakes did meaningfully harm mechanism performance. The experimental results persuaded the Wharton School to adopt the new mechanism and helped guide its practical implementation. It is hoped that the experimental design methodology may be of use to other market design researchers, either for evaluating or improving preference reporting for existing mechanisms or for bringing other new mechanisms that utilize rich preference information from theory to practice.

Eric Budish
Booth School of Business
University of Chicago
5807 South Woodlawn Avenue
Chicago, IL 60637
and NBER
eric.budish@chicagobooth.edu

Judd B. Kessler
The Wharton School
University of Pennsylvania
3620 Locust Walk
Philadelphia, PA 19104
and NBER
judd.kessler@wharton.upenn.edu

1 Introduction

The promise of market design research is that mechanisms designed using abstract microeconomic theory can be implemented in practice to solve real-world resource allocation problems. This promise has led to an explosion of research in matching and auction theory and has led to several well-known market design "success stories", in which a mechanism has made it all the way from theory to practice. These include auctions for wireless spectrum around the world and matching mechanisms for entry-level medical labor markets, public schools and organ transplantation.[1] To bring these market design mechanisms to practice often requires innovative academic research to help test the theory and evaluate its suitability for practice. In this spirit, this paper reports on a novel kind of laboratory experiment — based on bringing real market participants' real preferences into the laboratory, as opposed to endowing experimental subjects with artificial preferences as is typical in the market design experimental literature — that tested a new market design theory and helped shepherd it from theory to practice.[2]

The context is the problem of combinatorial assignment — matching bundles of indivisible objects to agents without the use of monetary transfers, e.g., matching students to schedules of classes — well known to be a difficult problem in market design. The theory literature on this problem contains mostly impossibility theorems that prove there is no perfect mechanism,[3] while the mechanisms used in practice have been shown to have critical flaws.[4] In an attempt to make progress on this problem, Budish (2011) proposed a new mechanism for combinatorial assignment, called approximate competitive equilibrium from equal incomes (A-CEEI). A-CEEI, unlike prior mechanisms, satisfies attractive properties of efficiency, fairness and incentives, though as the name implies only does so approximately.

At around the same time Budish (2011) was published, an opportunity to potentially implement a new mechanism arose at the Wharton School at the University of Pennsylvania. Wharton's mechanism, a fake-money auction used widely at many educational institutions,[5] was having the kinds of efficiency, fairness and incentives problems one would expect given

[1] On spectrum auctions, see Milgrom's (2004) and Klemperer's (2004) fittingly named books, "Putting Auction Theory to Work" and "Auctions: Theory and Practice", as well as Cramton, Shoham and Steinberg (2006), Ausubel, Cramton and Milgrom (2006), Levin and Skrzypacz (2016) and Milgrom and Segal (Forthcoming). On matching markets, see Roth's (2015*b*) book as well as Roth (2002, 2008), Roth and Peranson (1999), Abdulkadiroğlu and Sönmez (2003), Abdulkadiroğlu, Pathak and Roth (2005), Abdulkadiroğlu et al. (2005, 2006), Roth, Sönmez and Ünver (2004, 2005, 2007) and Abdulkadiroğlu, Agarwal and Pathak (2017).

[2] See Roth (2015*a*) for a survey of the literature on market design experiments as well as a detailed discussion of the present paper in Section 6.

[3] See Pápai (2001), Ehlers and Klaus (2003), Hatfield (2009) and Kojima (2009).

[4] See Sönmez and Ünver (2003, 2010), Krishna and Ünver (2008) and Budish and Cantillon (2012).

[5] See Sönmez and Ünver (2010) for a list of schools using this mechanism and a description of the (minor) design variations across institutions. See Section 2 for more details on Wharton's variant, which uses a fake-money Vickrey auction in an initial allocation round and then uses double auctions in subsequent rounds.

the theoretical criticisms of the mechanism (Sönmez and Ünver 2003, 2010) and the Wharton administration convened a committee to consider alternatives.

While attractive in theory, however, the A-CEEI mechanism makes an assumption that raises serious concern about its suitability for use in practice: "agents report their type". In Budish (2011), a student's type is an ordinal preference relation over all possible schedules of courses, much as in general-equilibrium theory a household's type is an ordinal preference relation over all possible consumption bundles. As is completely standard in mechanism design theory (Fudenberg and Tirole 1991, Myerson 1991, Bergemann and Morris 2005), agents are assumed to be able to simply "report their type" to the mechanism. But this assumption often strains reality, and A-CEEI is such a case. In a context such as Wharton's, there might be hundreds of millions of schedules in a given semester.

Clearly, in such settings, perfect preference reporting is an unrealistic goal, and whether market participants can report perfectly is an uninteresting question. Instead, the relevant question to answer before seriously considering bringing the theory to practice is whether market participants can report their preferences "accurately enough" to realize the benefits of the mechanism. Let us make this question more precise. In any practical implementation of the A-CEEI mechanism, participants cannot be expected to manually rank all schedules. Instead, participants must report a limited set of preference data — via what is known as a preference reporting language (Milgrom 2009, 2011) — that can be used to construct an ordinal ranking over schedules. The question is whether participants can report such preference data with sufficient accuracy — that is, whether the ordinal ranking generated by the preference data they report is close enough to the true preferences in their minds — that the efficiency and fairness benefits of A-CEEI are realized.

This positive question about A-CEEI's suitability in turn raises a deeper methodological question that pertains to market design more broadly. How can a researcher generate data that yields an assessment of preference reporting if agents' true preferences are fundamentally unknown? In the case of A-CEEI, how can we compare the ordinal ranking generated from the data agents report to the mechanism to agents' true preferences? How can we measure the extent to which inaccurate preference reporting harms mechanism performance?

We designed a novel kind of experiment to answer these questions. Before describing our experimental design, we first explain why a new kind of experimental design is needed. The traditional method used in market design experiments is to endow subjects with artificial preferences and offer monetary rewards based on how well the subjects perform in the mechanism as evaluated based on these preferences. For example, if in a multi-object matching experiment a subject is endowed with a value of \$25 for the bundle {A, B}, and then obtains the bundle {A, B} in the laboratory matching market, the subject would be compensated with a payment of \$25. While this technique has been extremely important

in the history of market design experiments and is invaluable for answering certain kinds of questions, it is a non-starter for our setting because it assumes away the central issue of the difficulty of reporting one's preferences.[6] If we endowed subjects with artificial preferences in a format that could be immediately reported to the mechanism, we would just be telling subjects their preferences and asking them to tell them right back to us, trivializing the reporting task.[7] If we endowed subjects with artificial preferences in a format different from what could be reported to the mechanism, this too misses the central question of interest. This latter exercise tests whether subjects can translate between preferences in one language the researcher created (for conveying preferences to the subject) and another language the researcher created (for reporting preferences to the mechanism). This does not test whether real market participants can translate their own real preferences — however these preferences are represented in their own minds — into data the mechanism can use.

The following analogy may prove useful. Imagine a market participant's mental representation of preferences is in English and data must be entered into the mechanism in Latin. As noted above, if we endowed preferences in Latin and asked the subject to report preferences in Latin, this would be trivial (it would simply be a transcription exercise). If, instead, we endowed preferences in some other language, say Greek, we could test whether the subject could translate Greek into Latin, but this is fundamentally different than the test of whether the subject can translate English to Latin. We cannot test our fundamental question with endowed preferences unless we could somehow endow preferences in English, but this would require us to know the structure of agents' mental representations of their preferences, which is fundamentally unknowable to the researcher.

Instead of endowing experimental subjects with artificial preferences, our experimental design brought real market participants' real preferences into the lab. Specifically, our experimental subjects were Wharton MBA students who were asked to report their real preferences over schedules of real Wharton courses using a realistic, professionally designed user interface. We then generated our primary data on preference reporting accuracy and

[6] In Roth's (2015*a*) recent survey of the literature on market design experiments, every laboratory experiment discussed uses the endowed preferences methodology with the exception of the present paper, which is discussed in detail in Section 6. Outside of market design, it is common to design laboratory experiments around participants' real preferences; famous examples include Kahneman, Knetsch and Thaler (1990) and Roth et al. (1991). Note that in these latter settings, theory testing is possible without endowing preferences: in dictator and ultimatum games, subjects' preferences are assumed to be known a priori (favoring more money to less), and in endowment effect experiments the quantity of trade is sufficient to establish the effect without knowing subjects' precise values for the objects. In market design experiments, in contrast, theory testing often requires the researcher to have precise knowledge of subjects' heterogeneous preferences, which the endowed preferences methodology directly produces.

[7] Given that the A-CEEI mechanism is approximately strategy-proof and we informed subjects as such (as Wharton did in practical implementation), this would simply be testing whether subjects believed the claim in the instructions that it is in their best interest to report their preferences truthfully. This is an interesting question in its own right (cf. Hassidim, Romm and Shorrer 2016, Li 2017, Rees-Jones 2018, Rees-Jones and Skowronek 2018) but not the question of interest here. See further discussion in Section 2.4.

mechanism performance using binary comparisons — questions of the form: Do you prefer Schedule A or Schedule B? The rationale behind the binary comparisons method is that while reporting preferences over all possible schedules via a preference reporting language is cognitively hard and likely to be inaccurate, reporting which one prefers between two specific schedules is cognitively simple and likely to be accurate. As will be described in detail in Section 2.5, our experiment used carefully tailored binary comparisons to generate data on preference reporting accuracy and to test whether subjects could report preferences accurately enough to realize the efficiency and fairness benefits of the mechanism. Comparing the performance of the mechanism with regard to efficiency and fairness measures based on binary comparisons to efficiency and fairness measures based on the reported preferences, we can quantify the harm caused by preference reporting mistakes.

In addition to allowing us to test the "agents report their type" assumption, there were two other advantages to using real market participants' real preferences. First, the realism enhanced the demonstration value of the experiment. Demonstration to policy makers who ultimately decide whether to implement a market design is a common goal of market design experiments (cf. Roth 2015*a*); using real market participants' real preferences yields a more realistic, and thus more persuasive, demonstration. Second, the realism facilitated a search for "side effects" of the mechanism; that is, issues left out of the theory that might be important for practice.[8] Issues left out of the theory are especially of concern here because A-CEEI had never been used before; many other market design implementations have had direct precedents that assuage these concerns.[9] Because our experimental subjects were real market participants who were playing in a realistic environment, we could search directly for side effects using surveys. The surveys, both quantitative and free-response, covered topics such as perceived fairness, satisfaction with received schedule, ease of use, transparency and overall "liking" of the mechanism.

An important disadvantage of our experimental approach is that subjects' behavior is not incentivized.[10] This lack of incentives likely caused subjects to exert less effort in the

[8]Our use of the term "side effects" is meant to analogize the FDA drug approval process. The first step in that process is not to test the efficacy of the drug (that is the last step), but rather to ensure that the drug is not harmful to humans for some unforeseen reason.

[9]In many other practical market design implementations, there were close precedents that could be used to convince practitioners that the theory worked as intended in practice; these precedents lessen the concern about unintended consequences of the theory. For example, the Gale-Shapley deferred acceptance algorithm was independently discovered and implemented by the medical profession in the 1940s, about 15 years before the publication of Gale and Shapley (1962). Roth and Peranson (1999) report on the successful modification of the Gale-Shapley algorithm to accommodate married couples. When the Gale-Shapley algorithm was implemented for school choice, the economists involved in the implementation could point to the algorithm's decades of success in the medical labor market. Doctors discovered the idea of pairwise kidney exchange in the late 1990s; the economists who became involved helped to optimize what had been an ad hoc process to increase the number of potential matches.

[10]A similar lack of incentives arises in market design studies that utilize surveys to elicit preferences and/or beliefs, such as Kapor, Neilson and Zimmerman (2018) and Rees-Jones (2018). See Bertrand and

laboratory than they would have if playing for real stakes, which in turn adds noise to subjects' behavior. We took care in the design to ensure that such noise pushes against finding accurate preference reporting and against our finding benefits of the A-CEEI mechanism (see Section 2.6 for a discussion).[11]

We briefly summarize the main results. Students reported their preferences accurately enough that A-CEEI outperformed the benchmark, the incumbent Wharton Auction, on each of our quantitative measures of efficiency and fairness, with most (though not all) differences statistically significant. The magnitudes were modest but all broadly consistent with the theory. However, we also found that subjects had significant difficulty with preference reporting (although large mistakes were comparatively rare) and that this difficulty meaningfully harmed mechanism performance. The efficiency and fairness improvement of A-CEEI over the Wharton Auction would have been substantially larger if not for preference reporting mistakes. The only negative side effect we found in the surveys was that students found A-CEEI to be somewhat of a "black box", i.e., non-transparent.

The experiment persuaded Wharton to adopt A-CEEI — implemented as "Course Match" beginning in Fall 2013 — and guided several aspects of its practical implementation.[12] Some limited data from the first year of implementation demonstrates that A-CEEI has increased equity in both total expenditure and the distribution of popular courses, and survey data suggest that A-CEEI has increased students' satisfaction with their assigned schedules, their perceptions of fairness and their overall satisfaction with the course allocation system. For example, the percentage of students responding that they found the course allocation mechanism "effective" or "very effective" increased from 24% in the last year of the Auction to 53% in the first year of A-CEEI, and the percentage of students who agreed or strongly agreed that the course allocation mechanism "allows for a fair allocation of classes" increased from 28% to 65%.

Our paper makes three contributions to the market design literature more broadly. First, the paper contributes to an ongoing dialogue in the literature about the importance of preference reporting and language design (cf. Milgrom 2009, 2011). We provide some of the first documented empirical evidence on the prevalence of preference reporting errors and the harm they can cause to a mechanism's performance (see also Hassidim, Romm and Shorrer 2016, Rees-Jones 2018, Rees-Jones and Skowronek 2018), while at the same time

Mullainathan (2001) for a general discussion of the benefits and costs of survey data.

[11] Also note that the lack of incentives is not intrinsically a feature of the experimental design methodology we propose (i.e., using real agents' real preferences and binary comparisons). If we could have offered with some probability that students would obtain in real life the schedule they obtained in the lab version of the mechanism, or a schedule they chose in a binary comparison, then all behavior would have been incentivized. However, we were unable to get the Wharton administration to provide such stakes in the laboratory experiment, for the obvious reasons.

[12] After Wharton elected to adopt the new mechanism in spring 2012 the work of practical implementation began in earnest. The engineering component of this work is reported in Budish et al. (2017).

showing that participants can report complex preferences accurately enough to realize the benefits of a mechanism with complex reporting requirements.

Second, the paper introduces a new experimental design methodology that allows researchers to evaluate market designs in the laboratory using real market participants' real preferences and appropriate binary comparisons. This methodology can be used to evaluate other market designs with non-trivial preference reporting requirements. This methodology may also be useful for evaluating decision supports for market designs, i.e., tools that are designed to help participants more accurately report their preferences. Such decision supports play an important role not only in market designs with complex preference reporting requirements such as A-CEEI, but also in settings where the preference reporting per se is simple but thinking through one's preferences is difficult, e.g., school choice (cf. Narita 2016, Kapor, Neilson and Zimmerman 2018). By comparing subjects' ability to report their preferences with and without a particular decision support, our methodology can identify the efficacy of that decision support and help optimize the performance of existing market designs.

Our method is a complement to the endowed preferences methodology, which has been at the heart of a rich experimental literature in market design.[13] Within matching, experiments using endowed preferences have explored decentralized markets (e.g., Echenique and Yariv 2013), including issues such as unraveling and congestion (e.g., Niederle and Roth 2009); the transition to centralized clearinghouses (e.g., Kagel and Roth 2000); and problems in those centralized clearinghouses such as strategic misreporting (e.g., Castillo and Dianat 2016, Echenique, Wilson and Yariv 2016) and clearinghouse collapse (e.g., McKinney, Niederle and Roth 2005). In addition, a rich line of experimental work has used endowed preferences to explore school choice mechanisms in the laboratory, including work comparing the performance of various mechanisms, such as deferred acceptance, the Boston mechanism, and top trading cycles (e.g., Chen and Sönmez 2006, Pais and Pintér 2008, Calsamiglia, Haeringer and Klijn 2010, Featherstone and Niederle 2016, Ding and Schotter 2017). Finally, endowed preference laboratory experiments have been used to explore new matching mechanisms (e.g., Fragiadakis and Troyan 2016, Hakimov and Kesten 2018) and to explore new incentives criteria for market design (e.g., Li 2017, Chen et al. 2018).

Third, our paper contributes a new theory-to-practice success story to the market design literature. This is valuable for two related reasons. First, market design implementations beget further market design implementations. The Wharton committee was already familiar

[13]Some of the earliest examples of experiments using the endowed preferences methodology include the early double auction experiments of Chamberlin (1948) and Smith (1962) and combinatorial auction experiments such as Rassenti, Smith and Bulfin (1982) and Goeree and Holt (2010). See Kagel, Lien and Milgrom (2010) for an interesting twist on the methodology that uses theory and simulations to guide which endowed preferences to explore.

with the work done by economists re-designing spectrum auctions and matching markets, and this gave the committee some comfort that economists might have something useful to say about their problem, too. Our specific market design implementation paves some new ground — the mechanism descends from general equilibrium theory as opposed to auction or matching theory, ordinary individuals are asked to report the kinds of complex preferences more commonly associated with high-stakes combinatorial auctions, and a lab experiment played a pivotal role in the adoption decision — so we have some hope that one day other researchers seeking to implement new market designs will be able to use our implementation as a helpful precedent, just as we used the spectrum auctions and matching markets as a helpful precedent.

The second reason, as emphasized by Roth (2002), is that academic work on the practical implementation of market design theory is an important complement to the theory itself. This work shows whether a particular theory is robust and raises new questions for theory to consider (e.g., the optimal design of preference reporting languages). As Roth (2002) writes: "Whether economists will often be in a position to give highly practical advice depends in part on whether we report what we learn, and what we do, in sufficient detail to allow scientific knowledge about design to accumulate. ... If the literature of design economics does mature in this way, it will also help shape and enrich the underlying economic theory."

The remainder of this paper is organized as follows. Section 2 describes the experimental design. Section 3 presents the results on fairness and efficiency. Section 4 analyzes preference reporting mistakes. Section 5 reports on the survey data and the search for unintended consequences of the mechanism. Section 6 reports on the first year of practical implementation and concludes.

2 Experimental Design

2.1 Real Market Participants' Real Preferences

Our experimental subjects were Wharton MBA students, recruited by an email sent by the Wharton administration (see Appendix A).[14] There were 132 subjects over eight experimental sessions, conducted in a computer lab at Wharton during the week of November 28, 2011 (see the full text of the experimental instructions in Appendix C).

[14]The email indicated that the study was voluntary but that participation was appreciated by the Dean's office and as a further inducement offered $250 to two randomly selected subjects per session. The email did not mention that the study was about course assignment. We wanted to attract student subjects who were generally representative of the Wharton MBA student body and to avoid attracting students who were disproportionally happy or unhappy with the current course auction. Subjects were statistically representative of the Wharton student population on every dimension except race and, importantly, were representative with regard to attitudes toward the Wharton Auction (see discussion in Section 2.6 and Table A1 in Appendix B).

Subjects were given a list of 25 Wharton course sections for the upcoming Spring 2012 semester. These courses were chosen by the Wharton Course Allocation Redesign Team (the "Wharton committee") to be representative of course offerings in the upcoming semester with a tilt towards popular courses (see the list of courses and sample descriptions in Appendix D). Each course section had a capacity of 3 to 5 seats.

Subjects were instructed that they would participate in two course allocation procedures, Wharton's current system and an alternative system, and that their goal in the study was to use each system to obtain the best course schedule they could given their own true preferences. Here is some of the key text from the experimental instructions:

> "While using each system, please imagine that it is the spring term of your second year at Wharton, so this will be your last chance to take Wharton classes. Please try to construct your most preferred schedule given the courses that are available."
>
> "In real life, we know you take these decisions very seriously. We ask that you take the decisions in this session seriously as well. We will provide you with time to think carefully while using each system."

We then gave subjects five minutes to look over the course offerings and think about their preferences before describing the first mechanism.

2.2 Flow of Each Experimental Session

In half of the sessions we ran the Auction first, and for half of the sessions we ran A-CEEI first.[15] Details of the mechanisms are in Sections 2.3 and 2.4, respectively. For each mechanism:

i. We read aloud the instructions for that specific mechanism.

ii. Subjects participated in that mechanism to assemble a schedule of spring 2012 courses (starting from a blank slate for each mechanism).

iii. Subjects responded to Likert-scale survey questions about their experience with the mechanism. See Section 5 for details of the surveys.

After subjects had participated in both mechanisms:

i. Subjects performed a series of binary comparisons between pairs of schedules. These binary comparisons were designed to provide measures of efficiency, fairness and preference reporting accuracy. See Section 2.5 for details of the binary comparisons.

[15] We did not find any significant differences in the results based on which mechanism was used first. See Appendix E for details of this analysis.

ii. Subjects responded to Likert-scale survey questions comparing the two mechanisms.

iii. Subjects provided free-form response comments.

2.3 Wharton Bidding Points Auction

At the time of the experiment, Wharton's Auction, a variant on the bidding points auction mechanism used at a wide variety of educational institutions (Sönmez and Ünver 2010), worked as follows. In the first round of the Auction, students would submit bids for courses, with the sum of their bids not to exceed their budget (of an artificial currency called bidding points). If a course had k seats, the k highest bidders for that course obtained a seat, and paid the $k+1^{st}$ highest bid. After this first bidding round there were then eight additional rounds, spaced over a period of time lasting from the end of one semester to the beginning of the next, in which students could both buy and sell courses using a double auction.[16]

Our laboratory implementation of the Wharton Auction was as similar as possible to the real Wharton Auction, subject to the constraints of the laboratory. For time considerations, we used four rounds instead of nine.[17] For the first round, subjects were given five minutes to select their bids, with an initial budget of 5,000 points. For the remaining three rounds, subjects were given two-and-a-half minutes to select their bids and asks. The experiment used the standard web interface of the real Wharton Auction so that it would be as familiar as possible to subjects. The instructions for the Auction were familiar as well, since all subjects had previously used the real Wharton Auction mechanism to pick their courses.

2.4 Approximate Competitive Equilibrium from Equal Incomes (A-CEEI)

A-CEEI has four steps: (i) students report their preferences, (ii) each student is assigned an equal budget (5,000 points in the experiment) plus a small random amount (used to break ties),[18] (iii) the computer finds (approximate) market-clearing prices, (iv) each student is

[16]While the first round of the auction closely resembles a real-money Vickrey auction, the attractive properties of the Vickrey auction do not translate to the fake-money setting. The mathematical difference is that preferences are not quasi-linear over objects and money because the money is fake and the game is finite. Intuitively, someone who bids 10,000 dollars in a real-money auction and loses to someone who bids 10,001 may be disappointed, but at least they can put their money to some alternative use, whereas a student who bids 10,000 points in a fake-money auction and loses to someone who bids 10,001 may end up graduating with a large budget of useless course-auction currency. As a result, unlike the Vickrey auction, the bidding points auction is not strategy-proof and equilibrium outcomes can be highly unfair and inefficient. Note, however, that if the game were infinitely repeated then unspent fake money would always have a future use and so the quasi-linearity assumption would be valid. See Prendergast (2017) for an implementation of a mechanism in this spirit in the context of allocating donated food to food banks across the US.

[17]In practice, the final allocation of popular courses (i.e., courses with a positive price) is mostly determined by the outcome of the first round. This gave the Wharton committee confidence that there would not be much lost by using four rounds instead of nine. In the lab, too, most of the action took place in the first round.

[18]Budish's (2011) result that prices exist for A-CEEI that (approximately) clear the market requires that students have non-identical budgets. See also Reny (2017) for a recent generalization of this result. The

allocated her most preferred affordable schedule — the affordable schedule she likes best given her report in step (i) based on her budget set in step (ii) and the prices found in step (iii).[19]

The instructions described the A-CEEI mechanism, which was unfamiliar to the subjects, and explained to subjects that their only responsibility in using the mechanism was to tell the computer their true preferences; the computer would then compute market-clearing prices and buy them the best schedule they could afford at those prices. Because our interest was in whether subjects could report their preferences accurately enough to realize the theoretical benefits of the A-CEEI mechanism — and not in testing whether subjects could infer the strategy-proofness of the mechanism — we explictly instructed subjects to be as truthful as possible in their preference reporting. The instructions advised students: "...you do not need to think about the prices of the courses or the values that other students assign to courses. You get the best schedule possible simply by telling the computer your true values for courses."[20] The instructions used the metaphor of providing instructions to someone shopping on your behalf to explain the rationale for reporting one's true preferences as

budgets can be arbitrarily close to equal but cannot be exactly equal. The intuition is that the budget inequality helps break ties. For example, suppose students A and B both place an extremely high value on course X, which has one available seat. If A's budget is 5000 and B's budget is 5001, then setting the price of course X to 5001 clears the market because B can afford it while A cannot. The Auction breaks ties in the auction itself rather than in the budgets. If both A and B bid 5000 points for course X, then the computer randomly selects one student to transact.

[19]See Budish (2011) for a more complete description of how A-CEEI works. See Othman, Budish and Sandholm (2010) and Budish et al. (2017) for the computer science behind how to calculate the market-clearing prices in step (iii).

[20]We thought seriously about whether or not to caveat our instructions by more specifically explaining that A-CEEI is only approximately strategy-proof, not exactly strategy-proof, and therefore there theoretically are conditions under which an agent could benefit from misreporting. For reasons outlined in detail here, we decided that the best advice we could provide subjects was to report their preferences truthfully, and that dwelling on the difference between approximate and exact strategy-proofness would be confusing. At any realized prices, truthful reporting is best because it ensures the student receives her most-preferred affordable bundle at those prices. For it to be profitable for a student to benefit from misreporting her preferences, it must be the case that the misreport advantageously influences prices while at the same time the misreport does not cause the student to get the wrong bundle at the influenced prices. Formally, by reporting preferences as u' instead of u, this changes prices from p to p', and the student gets more utility from the bundle the mechanism thinks she likes best at p' (based on her misreport u') than from the bundle she likes best at p (based on her true preferences u). The main reason why such misreports are hard to find, even in small markets, is that students require at most one unit of any particular course. Therefore, the "demand reduction" strategies that are typically used to profitably manipulate prices in multi-object allocation mechanisms do not work here: if a student reduces demand for a course this can indeed reduce the price for that course, but since reducing demand means pretending to want zero units instead of one unit, this does not do the student any good. A second reason why such misreports are likely to be hard to find is the black box nature of the approximate Kakutani fixed point computation. Footnote 31 of the 2010 working paper version of Budish (2011) gives an example of the kinds of profitable manipulations that were found in extensive computational exploration in small markets and they are non-intuitive. Since there is a risk to misreporting — one is no longer guaranteed one's most-preferred affordable schedule at the realized prices — and the benefits of misreporting are difficult, if not impossible, to realize, we decided the best advice we could give was to advise subjects to report truthfully. If either of the authors of this paper were participating in this market design, even in a small economy like the ones used in the laboratory, we would report truthfully.

accurately as possible. Here are some of the key excerpts:

> "Since the computer is going to optimally buy courses for you, your job is to provide the computer with all the information it needs about how much you value the courses. This is obviously very important, since the computer is going to buy the optimal schedule for you given only what it knows about how you value courses."

> "Another way to think about reporting your values to the computer is to imagine you are sending the computer to the supermarket with your food budget and a list of your preferences for ingredients for dinner. You want to report your true values so that the computer can make the right tradeoffs for you when it gets to the supermarket and observes the actual prices for each ingredient."

2.4.1 Preference Reporting Language

As emphasized in the Introduction, the theory behind A-CEEI makes the unrealistic assumption that agents can "report their type" — that is, an ordinal ranking over all feasible schedules — so that the mechanism can always select the agent's most-preferred affordable bundle from any possible choice set. In any practical implementation of A-CEEI, agents cannot be expected to directly report preferences over all possible bundles. Instead, agents will need to supply a more limited set of information that describes their preferences, using what is called a preference reporting language (cf. Milgrom 2009, 2011).

The preference reporting language we implemented in the lab, a simplified version of the language proposed in Othman, Budish and Sandholm (2010) and similar in spirit to the language proposed in Milgrom (2009), had two components. First, subjects could report cardinal item values, on a scale of 1 to 100, for any course section they were interested in taking; if they did not report a value for a course section its value was defaulted to 0.[21] Second, subjects could report "adjustments" for any pair of course sections. Adjustments assigned an additional value, either positive or negative, to schedules that had both course sections together. Adjustments are a simple way for students to express certain kinds of substitutabilities and complementarities.[22] Subjects did not need to report schedule constraints, which were already known by the system. The user-interface for this language, designed by Wharton information technology professionals, is displayed as Figure 1.

[21] We recommended reporting a positive value for at least 12 course sections to ensure receipt of a complete schedule of five courses.

[22] If subjects could report adjustments over arbitrary sets of courses rather than just pairs of courses, then in principle the language would allow students to express any possible ordinal ranking over schedules, making the language expressive as defined, e.g., in Nisan (2006). We explore limitations of the language in further detail in Section 4.

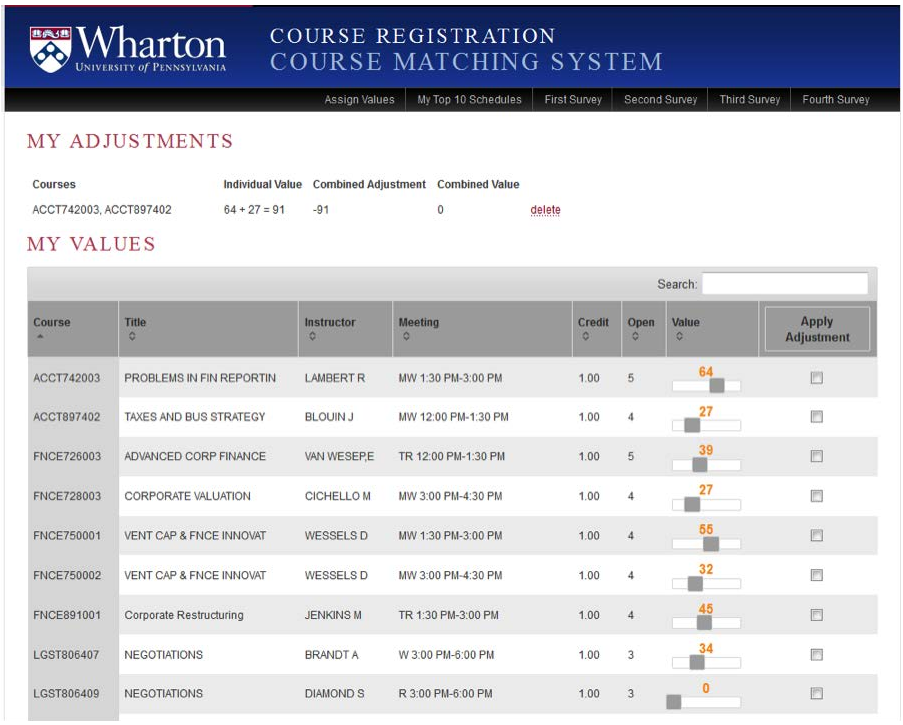
To calculate a subject's utility for a schedule, the system summed the subject's values for the individual courses in that schedule together with any adjustments (positive or negative) associated with pairs of courses in the schedule. The subject's rank order list over all schedules could thus be obtained by ordering schedules from highest to lowest utility.[23] Observe that this means that the cardinal preference information subjects submit for individual courses and pairs of courses induces an ordinal ranking over all feasible schedules, i.e., the language allows subjects to report a type.

We emphasize that while both we and the Wharton committee believed this preference reporting language to be reasonable — in particular, the Wharton committee felt strongly that adding more ways to express non-additive preferences would make the language too complicated — there is no reason to believe that this preference reporting language is optimal. As we discuss in the conclusion, optimal language design is an interesting open question for future research.

Given the complexity of preference reporting, and in particular the complexity of translating cardinal item values and adjustments into an ordering over schedules, we provided subjects with a decision support tool, the "top-ten widget" (see Figure 2), which allowed them to translate the preference information they had provided so far into a list of what the system currently calculated to be their 10 most-preferred schedules (displayed in order, with the accompanying sum of the cardinal utilities and adjustments next to each schedule). Subjects could use this widget at any time while reporting their values and could go back to make modifications to their values, e.g., if they realized the 10 schedules listed were not their favorites or were in the wrong order. Students were given 10 minutes to report their preferences.

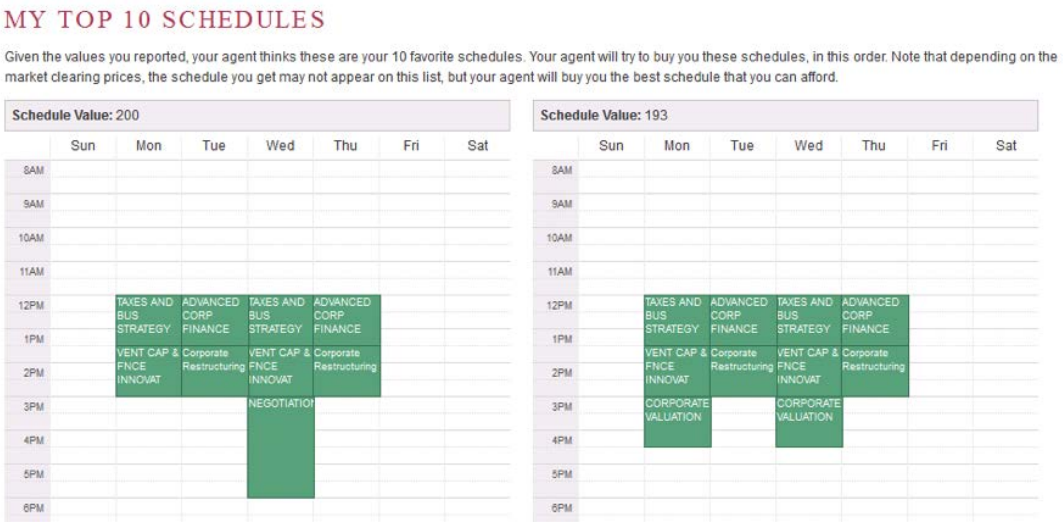
[23] Computationally, it is not necessary to ever formulate a student's complete rank order list over schedules. Instead, the question of what is a student's most-preferred affordable schedule at a given price vector can be translated into a mixed-integer program. This is an important computational advantage because integer programming, though NP-hard, is speedy in practice for problems of this size. The practical implementation of A-CEEI solves billions of integer programs in the process of finding approximate market clearing prices. See Budish et al. (2017) for more details on the computational procedure.

Figure 1: Screenshot of the A-CEEI User Interface



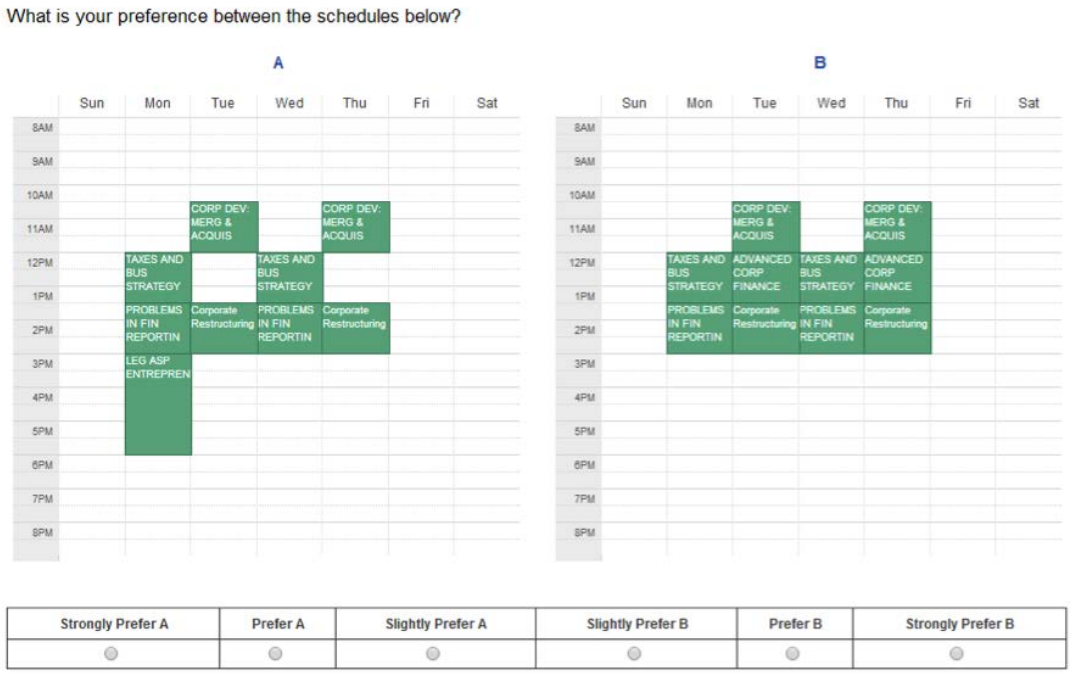
Notes: Figure 1 is a screenshot of the top of the user interface for preference reporting. Of the nine course sections that are visible, the hypothetical subject has reported positive values for the first eight. To make adjustments, subjects clicked two checkboxes in the far right column of the interface and were prompted to enter the adjustment in a dialog box. Any previously entered adjustments were listed at the top of the interface. The hypothetical subject has made one adjustment of -91, which tells the mechanism that getting the two accounting classes (i.e., the first two courses visible) together in his schedule together is worth 0, effectively reporting that the subject wants one or the other, but not both, accounting courses.

Figure 2: Screenshot of the Top-Ten Widget



Notes: Figure 2 is a screenshot of the top of the top-ten widget. It shows two feasible schedules of five courses each (e.g., "Taxes and Business Strategy" meets from 12:00-1:30 on Monday and Wednesday in both schedules) and the subject's reported utility for each of these schedules, listed as "Schedule Value". The rest of the top-ten schedules were shown below these, and subjects could scroll down the screen to see all 10.

Figure 3: Screenshot of a Binary Comparison Question



Notes: Figure 3 is a screenshot of a binary comparison. It shows two schedules and asks the subject to pick which of the two the subject prefers.

2.5 Binary comparisons

A simple methodological innovation, binary comparisons, is what allowed us to generate data on market design performance without knowing market participants' underlying true preferences. The logic behind the methodology is that while reporting one's type (i.e., ordinal preferences over every possible schedule) using the preference reporting language is cognitively complex and all but certain to be somewhat inaccurate, making a binary comparison between two specific schedules is cognitively simple and likely to accurately reflect true preferences.

More specifically, after using both mechanisms, subjects were shown up to 19 pairs of schedules, and asked to report which of the two schedules they preferred, on a scale of "Strongly Prefer", "Prefer" and "Slightly Prefer" for each schedule. See Figure 3 for a screenshot.

We designed the set of binary comparisons to yield data to test whether agents were able to report preferences accurately enough to realize the efficiency and fairness benefits of A-CEEI relative to the Auction as well as to provide data to directly test agents' preference reporting accuracy.

2.5.1 Efficiency and Fairness

Efficiency

Subjects' first and last binary comparisons were between the schedule the subject received under A-CEEI and the schedule she received under the Auction. This comparison was asked twice, as the first question and the last question, with the order of the schedules reversed.[24] These binary comparisons yield a simple social welfare comparison between the two mechanisms. Specifically, if more subjects prefer their A-CEEI schedule to their Auction schedule than vice versa, with similar strength of preference, this suggests that a social planner deciding between the two mechanisms should prefer A-CEEI, as should a student choosing between the two mechanisms from behind a veil of ignorance. Note that this comparison can be made at the individual-subject level, treating each subject as an independent observation for statistical tests, and also at the market-session level, aggregating up preferences to ask which of the two mechanisms generates more social welfare at the session level.[25]

[24] The schedule shown on the left in the first question was shown on the right in the last question. These binary comparisons were only asked if the schedules received under the two mechanisms were different.

[25] We are interested in both individual-level and session-level results and it is worth noting that there are inherent tradeoffs between the two. Looking at individual-level data reflects the fact that we care about individual agents being made better off by a mechanism and gives us more data to run our statistical tests, but ignores the session-structure of our data. Looking at session-level data respects the fact that mechanisms are, by definition, implemented at the market level, but gives us only eight sessions to run our statistical tests.

Ideally, we would also examine other measures of ex-ante social welfare. While we cannot use the binary comparisons alone to do so, we can use reported preferences data to calculate how utility (based on reported preferences) differs between the mechanisms. These exercises do not speak to our main question of whether agents can report their preferences accurately enough to realize the theoretical benefits of A-CEEI, but they provide additional evidence on the performance differences between the two mechanisms.[26]

Fairness

To measure fairness, each subject completed up to six binary comparisons per mechanism that directly assessed whether the subject envied another subject's schedule. Envy occurs when an individual prefers someone else's schedule to her own schedule; envy freeness is one of the oldest and most well established criteria of outcome fairness in economics (Foley 1967, Moulin 1995). To increase the chance of detecting envy, each subject was only shown schedules from the set of others' schedules that generated at least 50% of the utility of the subject's own A-CEEI schedule, based on the preferences the subject reported under A-CEEI. Restricting to this set aimed to ensure that subjects would face at least somewhat desirable alternative schedules when answering these binary comparisons. If more than six schedules of other subjects were in this set, six schedules from this set were chosen randomly by the computer to be used in binary comparisons. If six or fewer schedules were in this set, all schedules in the set were used in binary comparisons. This design choice makes the implicit assumption that schedules generating less than 50% of the utility of the subject's own A-CEEI schedule will not be envied, an assumption that we are able to evaluate ex post (see Appendix F).[27]

We use these binary comparisons and the definition of envy freeness to ask whether subjects experienced more envy under one mechanism than another. Similar to the analysis for efficiency, we will use these binary comparisons to generate a test of fairness at the individual-level (i.e., did a subject experience more envy under one mechanism than the other) and a test of fairness at the session level (i.e., did subjects in a market experience more envy under one mechanism than the other).

[26] We are also interested in measuring ex-post Pareto efficiency, both because it is a relevant efficiency property and because the theory in Budish (2011) shows that A-CEEI is approximately ex-post Pareto efficient. However, like additional measures of social welfare, it is infeasible to test for ex-post Pareto efficiency using binary comparisons. In particular, the number of potential Pareto-improving trades is combinatorially large, and includes both small trades (e.g., of one class for one class) and larger trades, but we have just a limited number of binary comparisons between complete schedules. We instead use our reported preference data to measure the number of Pareto-improving trades under the two mechanisms in Section 3.4. Again, however, this does not speak to our main question of whether agents can report their preferences accurately enough.

[27] Results in Appendix F show that while this assumption is unlikely to hold perfectly, its failure to hold works against us finding that A-CEEI generates less envy than the Auction.

Remark: Joint Tests

We emphasize that these binary comparison measures of efficiency and fairness are necessarily joint tests of preference reporting and the mechanisms. That is, these comparisons answer the question: is preference reporting accurate enough that A-CEEI is able to outperform the Auction on measures of efficiency and fairness? In addition, by comparing efficiency and fairness outcomes based on binary comparisons to the corresponding outcomes if we were to assume reported preferences were accurate, we can assess the extent to which imperfect preference reporting harmed mechanism performance.

2.5.2 Preference Reporting

All binary comparisons are tests of the A-CEEI preference reporting language, because we can assess whether the subject's binary choice between schedules is consistent with their preference report. In addition to the binary comparisons described above, we included five binary comparisons that were aimed specifically at preference reporting accuracy, which compared the schedule the subject realized under A-CEEI to the schedule that subject would have received (if distinct) if their budget had been 10% or 30% higher or 10% or 30% lower than it actually was. These binary comparisons provide local tests of preference reporting accuracy, examining schedules similar to the one the subject received. We investigate why subjects may have had difficulty reporting preferences in Section 4.

2.6 Discussion: Incentives

Before we present the results, we want to return to the issue discussed in the Introduction that decisions in our experiment are not incentivized. As described in detail in the Introduction, we could not use the endowed preferences methodology, since that would not allow us to test our fundamental research question of whether participants could report their real preferences accurately enough to realize the theoretical benefits of A-CEEI. As noted in footnote 11, we were not able to incentivize choices in our experiment, since doing so would have required giving subjects some positive probability of receiving — for a real upcoming Wharton spring semester — each of the schedules they constructed in the mechanisms and selected in the binary comparisons. The typical response when researchers are unable to offer desired incentives in a laboratory experiment is to attempt to run a field experiment. In the field, both real market participants' real preferences and incentives for their choices are usually already in place. Such an experiment might have randomly assigned students to use different course allocation mechanisms (e.g., assigned some to use A-CEEI and others

to use the Wharton Auction, each for a subset of the available spring semester seats).[28] Given the nature of the problem, however, running a field experiment was just as infeasible as providing incentives for our laboratory study.[29]

We therefore faced a design challenge. While we were able to bring real market participants' real preferences into a controlled laboratory environment, we were not able to incentivize their decisions and we needed to understand and mitigate any potential risk of the absence of incentives.[30]

The main risk is that subjects might not exert as much effort in an unincentivized experiment as they would in an incentivized one. We thus took care to design the experiement so that such lack of effort, if present in our setting, would bias against finding that agents could report their preferences accurately enough for A-CEEI to realize the benefits promised by the theory.

Imagine there are two kinds of experimental subjects, "triers" and "non-triers". Triers exert the same level of effort in the experimental tasks as they would if fully incentivized, while non-triers exert zero effort in the mechanisms and their binary comparison responses are pure noise, i.e., 50/50 coin flips. This noise from the non-triers biases towards less accurate preference reporting under A-CEEI and less ability to detect a difference in efficiency or fairness between A-CEEI and the Auction. This pushes against finding that subjects can report preferences accurately enough for A-CEEI to outperform the Auction: noise from the non-triers biases our results away from finding that subjects can report their preferences accurately and biases our results away from finding that A-CEEI improves efficiency and fairness relative to the Auction.

A subtler case is if the lack of incentives causes subjects to exert effort that is intermediate between full effort and pure noise. To understand what would happen in this case would require an understanding of the function mapping the level of effort to how well subjects perform in the experimental mechanisms and how accurately they reply to binary comparisons. We of course do not know this function, but, given that the Auction is familiar to subjects while A-CEEI is unfamiliar, we might expect partial effort to harm A-CEEI more

[28] To evaluate whether one mechanism outperformed the other, such a field experiment would preusmably also need an incentivized elicitation procedure, e.g., testing for envy by giving students the option to trade their realized schedule for the realized schedules of other students, with some positive probability.

[29] A field experiment was a non-starter at Wharton, presumably both for logistical reasons and due to concerns about students' perceptions of fairness. The prospect of such a field experiment also raises a Catch-22, since even if the Wharton Administration had considered such a field experiment, they would likely have wanted to see inital evidence that the mechanism could be successful — evidence of the kind generated by a laboratory experiment like ours.

[30] While subjects' decisions were not incentivized, subjects were compensated for their time in the form of two $250 prizes per session to randomly chosen subjects. The Wharton committee thought that two $250 prizes per session would be more appropriate and attractive compensation than paying each student the expected value of roughly $30. Suffice it to say, MBA students are different from the typical undergraduate subject pool.

than the Auction, which also pushes against finding that A-CEEI outperforms the Auction.

A second potential risk that is distinct from low effort, and which would bias some results in our hypothesized direction, is that students in the lab disliked the Wharton Auction in practice and thus attempted to sabotage its performance in the lab. While we cannot rule out this possibility entirely (nor could we even if the experiment were incentivized), a few things give us comfort. First, the subjects in the experiment were representative of the Wharton student body as a whole, both on demographic measures and, crucially, on their perception of the Wharton Auction's effectiveness (see Appendix Table A1 in Appendix B).[31] Second, subjects were recruited to the experimental sessions by an email that came from the Wharton administration that did not mention course allocation[32] and subjects were explicitly asked in the experimental instructions to take their decisions seriously in the lab just like they do in real life. Our impression, given the attentiveness of the subjects and the questions they asked during the sessions, is that the Wharton students in the laboratory took this direction seriously.

3 Results on Fairness and Efficiency

As described in Section 2, we are interested in assessing whether agents are able to report their preferences accurately enough to realize the efficiency and fairness benefits promised by A-CEEI and to assess the harm caused by imperfect preference reporting.

Our main results comparing A-CEEI to the Auction appear in the two-by-two-by-two matrix labeled Table 1, which presents results of our efficiency tests (top panel) and fairness tests (bottom panel). Given that experimental subjects participate in the market with the other subjects in their session, the table presents results at the individual-subject level (left column) and the market-session level (right column).

We provide our main tests of whether subjects can report their preferences accurately enough for A-CEEI to outperform the Auction using binary comparison data (first row of each panel) and give an indication of the extent to which imperfect preference reporting harmed mechanism performance by showing the same tests using reported preference data (second row of each panel). The difference in these tests gives a sense of magnitudes for the harm caused by preference reporting mistakes (a statistical test of this difference is in the bottom row of each panel). The tests using reported preference data also provide an upper bound on the performance benefits of A-CEEI relative to the Auction if preference

[31] We used anonymous Wharton IDs to match experimental subjects to data from an administration survey conducted at the end of each school year. Our laboratory subjects rated the Wharton Auction's "effectiveness" an average of 4.69 on a scale of 0 to 7, essentially identical to the overall Wharton average of 4.68.

[32] The recruitment email (as shown in Appendix A) did not mention course allocation to help ensure we did not attract people with particularly strong views on the existing auction.

reporting could be made more accurate, e.g., through education and training of students or by giving students more time to think about and report their preferences than was possible in the laboratory.

We will discuss the results from the top panel on efficiency in Section 3.1 and the bottom panel on fairness in Section 3.2. To complement the results presented in Table 1, in Section 3.3, we present robustness tests of our binary comparison results that utilize the rich nature of the binary comparison data (e.g., including the intensity of preference) and otherwise redefine our outcome variables. These robustness tests show that our results are similar under different definitions of our key outcome variables. In Section 3.4, we use reported preference data to further explore performance differences of the two mechanisms, abstracting away from preference reporting mistakes.

We make two remarks regarding methodology. First, we believe it is appropriate to use one-sided statistical tests for the analyses in this section. In the tests based on reported preferences (cells (C), (D), (G) and (H)), we are testing directional predictions based on the theoretical efficiency and fairness benefits of A-CEEI and the theoretical efficiency and fairness problems of the Auction (Sönmez and Ünver 2010). In the tests based on binary comparisons (cells (A), (B), (E) and (F)), a one-sided test is appropriate given the nature of our research question. If we reject the null, we will conclude that subjects were indeed able to report their preferences accurately enough to realize the theoretical efficiency and fairness benefits of A-CEEI. If we fail to reject the null, we will conclude that subjects had sufficient difficulty with preference reporting that the theoretical benefits failed to manifest.[33] That said, we recognize that some readers may prefer two-sided tests; two-sided tests would double all p-values in the table, and in particular would cause the individual-subject binary comparison result to go from marginally signficant at the 10 percent level to insignificant.

Second, while we report statistical tests separately for each of the eight cells in the matrix, we consider the gestalt of the results as more informative than any individual test. More specifically, we take comfort that all of the binary comparison results are in the same direction, and that the binary comparison and reported preference results are all consistent with the conclusion that subjects reported accurately enough to realize the theoretical benefits of A-CEEI but that imperfect preference reporting harmed mechanism performance.

[33]Note that even if the Auction were to perform much better than A-CEEI, we would *not* conclude that the Auction is a better mechanism on efficiency or fairness grounds. Rather, we would go back to the drawing board regarding the preference reporting language.

Table 1: Efficiency and Fairness

		Aggregation Level	
Outcome	**Data**	*Individual-Subject*	*Market-Session*
Efficiency	*Binary Comparison*	(A) 56 - Prefer A-CEEI 42 - Prefer Auction 17 - Identical outcomes 17 - Indeterminate preference $p = 0.094$	(B) 6 - Prefer A-CEEI 0 - Prefer Auction 2 - Tie $p = 0.016$
	Reported Preference	(C) 79 - Prefer A-CEEI 35 - Prefer Auction 17 - Identical outcomes 1 - Indeterminate preference $p < 0.001$	(D) 7 - Prefer A-CEEI 0 - Prefer Auction 1 - Tie $p = 0.008$
	Both	*Test that Binary Comparison and Reported Preference classifications are the same:* $p < 0.001$	 $p = 0.500$
Fairness	*Binary Comparison*	(E) 40 - Less Envy A-CEEI 23 - Less Envy Auction 65 - No Envy either 4 - Same Envy both $p = 0.022$	(F) 5 - Less Envy A-CEEI 1 - Less Envy Auction 2 - Tie $p = 0.109$
	Reported Preference	(G) 35 - Less Envy A-CEEI 4 - Less Envy Auction 93 - No Envy either 0 - Same Envy both $p < 0.001$	(H) 8 - Less Envy A-CEEI 0 - Less Envy Auction 0 - Tie $p = 0.004$
	Both	*Test that Binary Comparison and Reported Preference classifications are the same:* $p = 0.072$	 $p = 0.125$

Notes: See definitions for the labels listed in the table in the sections of the main text corresponding to each cell (A)-(H). For *Efficiency* (top panel), we test whether agents are more likely to prefer their A-CEEI schedule to their Auction schedule. For *Fairness* (bottom panel), we test whether subjects experience less envy in A-CEEI than in the Auction. P-values reported in cells (A)-(H) are one-sided sign tests. P-values reported in the "Both" rows are matched-pair sign tests that compare a subject's (or session's) classification based on binary comparisons to that subject's (or session's) classification based on reported preferences, with the null hypothesis that the median of these differences is equal to 0.

3.1 Efficiency Tests

Binary Comparison, Individual-Subject (Table 1, Cell A)

As described in Section 2.5, our binary comparisons on efficiency provide a measure of social welfare by asking subjects which of the two mechanisms they prefer based on their realized schedules. In particular, we asked subjects who received different schedules from the two mechanisms whether they preferred the schedule they received under A-CEEI or the schedule they received under the Auction. This question was asked twice, once as the first binary comparison and once as the last binary comparison with the order of the schedules reversed between the two.

Consequently, individual subjects can fall into one of four mutually exclusive groups based on their binary comparison data. Subjects can either: prefer their A-CEEI schedule in both binary comparisons (which we label "Prefer A-CEEI"), prefer their Auction schedule in both binary comparisons ("Prefer Auction"), not display a consistent preference between the two schedules they received ("Indeterminate preference") or receive the same schedule from both mechanisms ("Identical outcome").[34]

As reported in Cell A of Table 1: 56 subjects Prefer A-CEEI, 42 subjects Prefer Auction, 17 subjects have an Indeterminate preference and 17 subjects receive Identical outcomes. To test whether A-CEEI outperforms the Auction, we treat each subject as an independent observation, assign subjects with an indeterminate preference or identical outcomes as having no preference between the mechanisms and perform a one-sided sign test.[35] The test yields a p-value of $p = 0.094$. This result suggests that subjects are able to report their preferences accurately enough for A-CEEI to outperform the Auction on this efficiency measure, though only at the 10% significance level.

Binary Comparison, Market-Session (Table 1, Cell B)

To conduct our session-level tests, we aggregate these individual preferences up to the session level based on a majority-rule social welfare criterion. We count the number of Prefer A-

[34] As shown in Figure 3, subjects were not given an option to report that they were indifferent between two schedules and so seeming preference reversals among subjects with an Indeterminate preference may be a reflection that some subjects felt indifferent between the two schedules. It could also be an indication of subject errors or random choices. As discussed in Section 2.6, the extent to which subjects respond randomly works against us finding any differences between the mechanisms.

[35] The sign test is a standard non-parametric test. We treat "Prefer A-CEEI" (and, later, "Less Envy A-CEEI") as A-CEEI outperforming the Auction, "Prefer Auction" (and, later, "Less Envy Auction") as the Auction outperforming A-CEEI and all other classifications as A-CEEI and the Auction performing equally well. The sign test assigns a positive value to an observation in which A-CEEI outperforms the Auction and a negative value to an observation in which the Auction outperforms A-CEEI. It then tests whether the median of these values is equal to 0. Note that with data of this form, the sign test is equivalent to a binomial probability test that tests whether our data could have come from a data generating process in which A-CEEI outperforms the Auction and the Auction outperforms A-CEEI are equally likely to arise.

CEEI and Prefer Auction in each session. If there are more of the former, we classify the session as "Prefer A-CEEI"; if there are more of the latter, we classify the session as "Prefer Auction"; and if there are an equal number, we classify the session as a "Tie".

As reported in Cell B of Table 1: 6 sessions Prefer A-CEEI, 0 sessions Prefer Auction and 2 sessions are a Tie. To test whether A-CEEI outperforms the Auction, we treat each session as an independent observation and perform a one-sided sign test. The test yields a p-value of $p = 0.016$. Looking at the market-session level reaffirms the individual-subject level results and indicates that agents are able to report their preferences accurately enough for A-CEEI to outperform the Auction on this efficiency measure.

Reported Preference, Individual-Subject (Table 1, Cell C)

The second row of Table 1 runs the same tests as the row above, but uses reported preference data rather than binary comparison data. Notice that we still have the same four classifications as when analyzing the binary comparison data in Cell A, but definitions have changed slightly since preferences are based on reported preference data. Subjects' preference reports may imply they receive higher utility from their A-CEEI schedule than their Auction schedule (which we label Prefer A-CEEI), receive higher utility from their Auction schedule than their A-CEEI schedule (Prefer Auction) or receive the same utility from different schedules from each of the two mechanisms (Indeterminate preference). If they receive the same schedule from both mechanisms we again use the label Identical outcome.

As reported in Cell C of Table 1: 79 subjects Prefer A-CEEI, 35 subjects Prefer Auction, 1 subject has an Indeterminate preference and 17 subjects receive Identical outcomes. A one-sided sign test yields a p-value of $p < 0.001$.

Reported Preference, Market-Session (Table 1, Cell D)

Applying the same majority-rule social welfare criterion to the individual preferences based on reported preferences yields a test of whether A-CEEI outperforms the Auction at the market-session level based on reported preferences. As reported in Cell D of Table 1: 7 sessions Prefer A-CEEI, 0 sessions Prefer the Auction and 1 session is a Tie. A one-sided sign test yields a p-value of $p = 0.008$.

Discussion

Results from the top row of Table 1 demonstrate that subjects are able to report preferences accurately enough to realize the efficiency benefits of A-CEEI. At both the individual level ($p = 0.094$) and the session level ($p = 0.016$), A-CEEI schedules are preferred to Auction

schedules. In addition, reported preference data suggests that absent preference-reporting mistakes A-CEEI would dramatically outperform the Auction.

Comparing results in Cells A and C allows us to test whether A-CEEI outperforms the Auction to a statistically significantly greater extent in reported preferences data than in binary comparison data. For the individual-subject data, we run a matched-pair sign test that compares a subject's classification based on binary comparisons to that subject's classification based on reported preferences, with the null hypothesis that the median of these differences is equal to 0. As shown in the bottom row of the efficiency panel, this test yields a p-value of $p < 0.001$. This suggests that while subjects report their preferences accurately enough for A-CEEI to outperform the Auction by a slim margin using individual-subject data, preference-reporting mistakes significantly harmed mechanism performance. Note that we do not see a significant difference when we run a similar matched-pair sign test on the session-level data.

3.2 Fairness Tests

Binary Comparison, Individual-Subject (Table 1, Cell E)

As described in Section 2.5, our binary comparisons on fairness allow us to investigate whether subjects experience less envy in A-CEEI than in the Auction. In particular, for each mechanism, each subject was asked to compare her realized schedule from that mechanism to (up to six) desirable schedules that other subjects in her session received from that mechanism. This generates a measure of how many schedules each subject envies in each mechanism.

Consequently, individual subjects can again be classified into one of four mutually exclusive groups based on their binary comparison data. Subjects can either: experience less envy under A-CEEI than the Auction (which we label "Less Envy A-CEEI"), experience less envy under the Auction than A-CEEI ("Less Envy Auction"), experience no envy under either mechanism ("No Envy either") or experience the same amount of envy (i.e., envy the same positive number of others' schedules) in both mechanisms ("Same Envy both").

As reported in Cell E of Table 1: 40 subjects are classified as Less Envy A-CEEI, 23 subjects are Less Envy Auction, 65 subjects are No Envy either and 4 subjects are Same Envy both.[36] To test whether A-CEEI outperforms the Auction, we treat each subject as

[36]These 65 subjects classified as No Envy either include six subjects for whom we do not collect data on envy due to a bug in our survey code: in the first three sessions we did not collect binary comparison data from subjects who received the same schedule under both A-CEEI and the Auction. While this bug was unfortunate, we believe, if anything, it is likely to work against us finding less envy under A-CEEI than the Auction. We come to this conclusion by looking at the other 11 subjects with identical A-CEEI and Auction schedules. Among this group, nine are No Envy either and two are Less Envy A-CEEI. Consequently, if the missing six subjects were similar to these 11, their data would have made our results weakly stronger.

an independent observation and perform a one-sided sign test. The test yields a p-value of $p = 0.022$. This finding demonstrates that subjects experience less envy under A-CEEI than under the Auction.

Binary Comparison, Market-Session (Table 1, Cell F)

As above, we compute our session-level results by aggregating up the individual classification as described with regard to Cell E to the session level. We count the number of Less Envy A-CEEI and Less Envy Auction in each session. If there are more of the former, we classify the session as "Less Envy A-CEEI"; if there are more of the latter, we classify the session as "Less Envy Auction"; and if there are an equal number, we classify the session as a "Tie".

As reported in Cell F of Table 1: 5 sessions are Less Envy A-CEEI, 1 is Less Envy Auction and 2 are a Tie. To test whether A-CEEI outperforms the Auction we treat each session as an independent observation and perform a one-sided sign test. The test yields a p-value of $p = 0.109$. Consequently, while we find statistically significant results with regard to fairness at the individual level, we have only directional evidence in support of A-CEEI outperforming the Auction at the session level.

Reported Preference, Individual-Subject (Table 1, Cell G)

Again, the second row of the fairness panel runs the same tests as the row above, but uses reported preference data rather than binary comparison data. We focus on the same subjects and the same comparison schedules but measure envy based on whether a subject's reported preferences suggest they get more utility from another subject's schedule than their own schedule. We then generate the same four classifications as when analyzing the binary comparison data.

As reported in Cell G of Table 1: 35 subjects are Less Envy A-CEEI, 4 subjects are Less Envy Auction, 93 subjects are No Envy either and 0 subjects are Same Envy both. A one-sided sign test yields a p-value of $p < 0.001$.

Reported Preference, Market-Session (Table 1, Cell H)

As above, we classify sessions based on the number of subjects in each session with the individual classifications in Cell G. As reported in Cell H, this exercise finds that all eight of the sessions are classified as Less Envy A-CEEI. A one-sided sign test yields a p-value of $p = 0.004$.

Discussion

Results from the binary choice data show that subjects are able to report the preferences accurately enough to realize the fairness benefits of A-CEEI. At the individual level, we find that subjects are less likely to experience envy under A-CEEI than under the Auction ($p = 0.022$). At the session level, the pattern of results is directionally consistent but not statistically significant ($p = 0.109$). Again, as expected, A-CEEI dramatically outperforms the Auction when abstracting away from preference reporting mistakes.

Comparing results in Cells E and G allows us to test whether A-CEEI outperforms the Auction to a statistically significantly greater extent in reported preferences data than in binary comparison data. Again we run a matched-pair sign test that compares a subject's classification under binary comparison data to that subject's classification under the reported preference data. As reported in the bottom row of the fairness panel, this test yields a p-value of $p = 0.072$. This suggests that while subjects report their preferences accurately enough for A-CEEI to outperform the Auction, preference-reporting mistakes marginally statistically significantly harmed mechanism performance. We find a similar, directional result at the session level (one-sided sign test, $p = 0.125$).

3.3 Robustness

There are a number of ways one might consider performing robustness tests on the binary comparison results above. Table 2 shows a variety of such tests. The top panel again focuses on efficiency and the bottom panel focuses on fairness.

Starting with efficiency, our first approach (Row 1 of Table 2) is to make our definition of preference stricter. Under this definition, subjects are only classified as Prefer A-CEEI if they state that they Prefer or Strongly Prefer their A-CEEI schedule to their Auction schedule in both binary comparisons (and likewise for the Auction). Subjects who state that they Slightly Prefer their favored schedule in either binary comparison are now classified as Indeterminate preference. Under this stricter definition, fewer subjects are classified as having a preference, but the results comparing A-CEEI to the Auction still suggest that A-CEEI is preferred to the Auction, at least marginally statistically significantly (see Cell I, $p = 0.057$). We get similar, directional results aggregating this new measure up to the session level (see Cell J, $p = 0.109$).

Table 2: Binary Comparison Robustness

		Aggregation Level	
Outcome	**Data**	*Individual-Subject*	*Market-Session*
Efficiency	(1) *"Prefer" or "Strongly Prefer"*	(I) 47 - Prefer A-CEEI 32 - Prefer Auction 17 - Identical outcomes 36 - Indeterminate preference $p = 0.057$	(J) 5 - Prefer A-CEEI 1 - Prefer Auction 2 - Tie $p = 0.109$
	(2) *Average Intensity*	(K) 59 - Prefer A-CEEI 47 - Prefer Auction 17 - Identical outcomes 9 - Indeterminate preference $p = 0.143$	(L) 5 - Prefer A-CEEI 1 - Prefer Auction 2 - Tie $p = 0.109$
Fairness	(3) *"Prefer" or "Strongly Prefer"*	(M) 36 - Less Envy A-CEEI 14 - Less Envy Auction 80 - No Envy either 2 - Same Envy both $p = 0.001$	(N) 6 - Less Envy A-CEEI 0 - Less Envy Auction 2 - Tie $p = 0.016$
	(4) *Binary Envy*	(O) 31 - Less Envy A-CEEI 17 - Less Envy Auction 65 - No Envy either 19 - Same Envy both $p = 0.030$	(P) 5 - Less Envy A-CEEI 2 - Less Envy Auction 1 - Tie $p = 0.227$
	(5) *Binary envy "Prefer" or "Strongly Prefer"*	(Q) 28 - Less Envy A-CEEI 11 - Less Envy Auction 80 - No Envy either 13 - Same Envy both $p = 0.005$	(R) 5 - Less Envy A-CEEI 2 - Less Envy Auction 1 - Tie $p = 0.227$

Notes: See definitions for the labels listed in rows (1)-(5) in the text of Section 3.3. P-values reported in each cell are one-sided sign tests.

Our second approach (Row 2 of Table 2) is to use the average intensity of subjects' preferences across the two binary comparisons, which allows us to assign a preference to eight additional subjects who previously were classified as having an Indeterminate preference because they reported that they preferred their A-CEEI schedule in one of the two binary comparisons and the Auction schedule in the other, with differing intensities. In this robustness test, we assign them a preference for A-CEEI if they indicated a stronger preference when they said they preferred their A-CEEI schedule than when they said they preferred their Auction schedule (three subjects) and assign them a preference for the Auction if the

opposite (five subjects). This yields an overall count of 59 subjects preferring A-CEEI and 47 preferring the Auction (see Cell K, $p = 0.143$). We again get similar directional results at the session level (see Cell L, $p = 0.109$). Taken together, our robustness tests are qualitatively similar to the main results presented in Table 1, although these additional tests provide slightly less statistical confidence.

Turning to fairness, we report on three robustness tests. Our first approach (Row 3 of Table 2) is again to use a stricter definition of preference, which treats subjects as envying another subject's schedule only if they Prefer or Strongly Prefer the other subject's schedule to their own. This allows us to classify fewer subjects as Less Envy A-CEEI and Less Envy Auction but our results remain strong with subjects experiencing less envy under A-CEEI than the Auction at both the individual level (see Cell M, $p = 0.001$) and the session level (see Cell N, $p = 0.016$).

Our second approach (Row 4 of Table 2) is to consider envy freeness as a 0-1 criterion and ask whether subjects envy at least one other subject's schedule in each mechanism. We now classify subjects as having Less Envy A-CEEI if they do not experience any envy in A-CEEI but do experience envy in the Auction, and Less Envy Auction if they do experience envy under A-CEEI but do not experience any envy under the Auction. We classify fewer subjects as Less Envy A-CEEI and Less Envy Auction, because now we treat anyone who experiences envy under both mechanisms as Same Envy both, even if the number of schedules they envy is different across the two mechanisms. However, our results remain statistically significant with subjects being less likely to experience envy under A-CEEI than the Auction (see Cell O, $p = 0.030$). We get similar, directional results at the session level (see Cell P, $p = 0.227$).

Our third approach (Row 5 of Table 2) combines the two previous approaches, using the binary measure of envy freeness as in Row 4 but using the stricter envy definition as in Row 3. Our results remain significant in this row with subjects experiencing less envy in A-CEEI than the Auction at the individual level (see Cell Q, $p = 0.005$). We again get similar, directional results at the session level (see Cell R, $p = 0.227$).

Taken together, these tests show that our results are robust, and different definitions of our outcome measures yield qualitatively similar results. While the level of statistical significance differs according to the test, the general pattern does not.

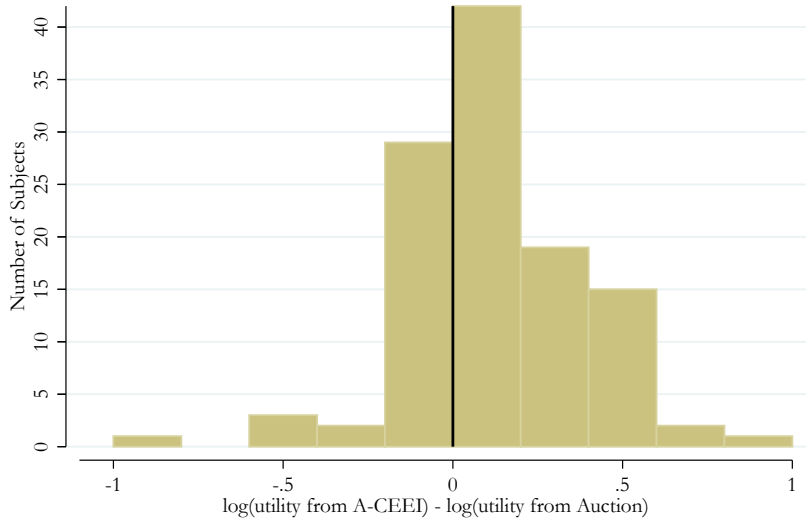
3.4 Additional Efficiency Analyses Based on Reported Preferences

As raised in Section 2.5, the reported preference data allows us to analyze additional measures of ex-ante social welfare and ex-post Pareto efficiency. These results, rather than speaking to our main question of whether agents can report their preferences accurately enough to realize the theoretical benefits of A-CEEI, instead explore how A-CEEI com-

pares to the Auction under the assumption of perfect preference reporting. These analyses can be interpreted as providing a further sense of magnitudes for the upper bound on the performance of A-CEEI relative to the Auction.

First, we can look directly at the differences in utility between the schedules a subject received from the two mechanisms. For each subject, we calculate *log(utility from A-CEEI) - log(utility from Auction)*, where "utility from A-CEEI" is the cardinal utility generated by the schedule the subject received under A-CEEI and "utility from Auction" is the cardinal utility generated by the schedule the subject received under the Auction. A histogram of these differences for the 114 subjects who get different utility from their two realized schedules is shown below as Figure 4.

Figure 4: Distribution of *log(utility from A-CEEI) – log(utility from Auction)*



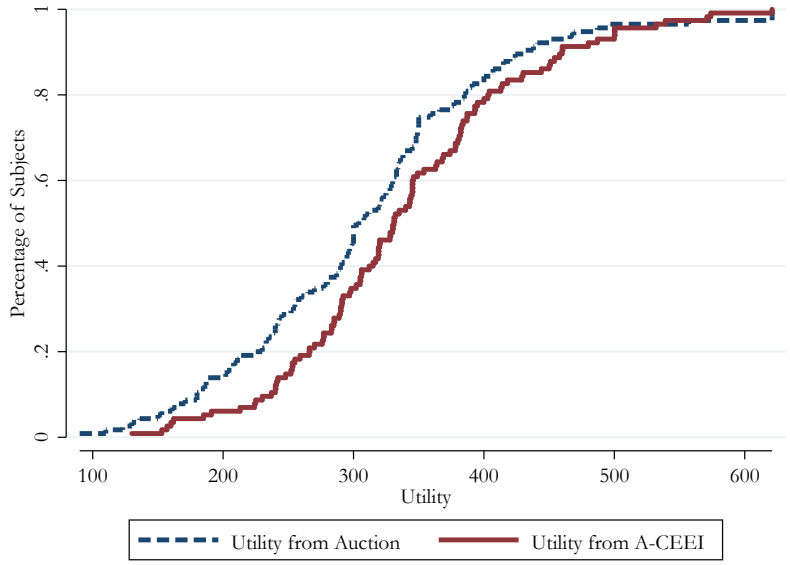
Notes: Figure 4 shows a histogram of the difference in *log(utility from A-CEEI) - log(utility from Auction)*. The graph excludes the 18 subjects who got the same utility from both schedules. One observation had difference in *log*(*utility*) more negative than -1 and is included in the furthest left bar of the histogram.

The majority of the mass of the histogram is to the right of 0 in Figure 4, a visual confirmation of the fact that, based on reported preferences, 69% (79/114, see Cell C of Table 1) of subjects who get different utility from their two realized schedules prefer their A-CEEI schedule to their Auction schedule. Moreover, the winners win more than the losers lose; 37 students have at least a 20% utility increase when comparing the Auction to A-CEEI, whereas only six students have at least a 20% utility decrease when comparing the Auction to A-CEEI.

Second, we can compare the distribution of utilities from realized schedules coming from each mechanism. Figure 5 plots these distributions for the same 114 subjects analyzed in

Figure 4. The distribution of utilities under A-CEEI second-order stochastically dominates the distribution under the Auction. This implies that a utilitarian social planner prefers the distribution of outcomes under A-CEEI to that under the Auction, so long as the planner has a weak preference for equality (the social welfare analogue of risk-aversion). However, the right tail of outcomes from the Auction generates higher utilities than the right tail of outcomes from A-CEEI. This arises since some people "win" the Auction, achieving schedules that are quite desirable and unattainable under A-CEEI. Consequently, we do not obtain first-order stochastic dominance.

Figure 5: Distribution of Utility from A-CEEI and the Auction



Notes: Figure 5 plots the CDF of cardinal utility based on subjects' reported preferences for both the Auction and A-CEEI. Three utilities (two in the Auction and one in A-CEEI) are above 2,000 and have been Winsorized at 621, the next-highest utility value (roughly the 99^{th} percentile).

Third, we examine ex-post Pareto efficiency. We formulate an integer program that solves for the maximum number of Pareto-improving trades in each session given subjects' reported preferences and the initial allocation arrived at in the experiment.[37] The theory of A-CEEI shows that it is approximately ex-post Pareto efficient. However, there may be Pareto-improving trades because of the small amount of market-clearing error that is sometimes necessary to run the mechanism.[38] The Auction is not Pareto efficient even

[37]We restrict attention to trades in which each subject in the trade gives and gets a single course seat. A subject may engage in an unlimited number of trades, and a trade may involve arbitrarily many subjects. An additional fictitious player called the "registrar" holds all unused capacity and has zero utility from each course.

[38]Budish (2011) shows that there need not exist prices that exactly clear the market, but guarantees existence of prices that clear the market to within a small amount of approximation error. See Reny (2017) for a recent generalization. In the theory, error is defined as the square root of the sum of squares of excess

approximately (see Sönmez and Ünver 2010). Table 3 reports the results of this exercise. As predicted by the theory, there is substantially less scope for Pareto-improving trades under A-CEEI than under the Auction.

Table 3: Ex-Post Pareto Efficiency

	Auction	A-CEEI	Test of proportions (one-sided)
# of Pareto-improving trades detected (% of course seats)	251 (31.7%)	44 (5.6%)	$p < 0.001$
# of students involved in at least one trade (% of students)	95 (72.0%)	23 (17.4%)	$p < 0.001$

Notes: Table 3 reports the results of an integer program that solves for the maximum number of Pareto-improving trades in each session based on subjects' reported preferences.

4 Difficulty with Preference Reporting

The efficiency and fairness results in Section 3 showed that difficulty with preference reporting meaningfully harmed mechanism performance. While A-CEEI did outperform the Auction in our efficiency and fairness tests based on the binary comparisons data, which reflect the difficulty of preference reporting, the outperformance was economically much larger in our measures based on the reported preferences data, which assume that preference reporting is perfect. In this section, we use data from our experiment to explore subjects' ability to report their preferences. Our goals are both to understand the causes of difficulty with preference reporting and to identify ways that preference reporting accuracy might be improved.

Conceptually, we distinguish between two possible reasons why subjects' preference reports might not reflect their underlying true preferences, i.e., why the "agents report their type" assumption might fail. First, subjects may have had difficulty using the preference reporting language we provided in the lab to express their underlying true preferences, even though in principle it was mathematically feasible for them to do so with the language. Second, there are some kinds of preferences that mathematically cannot be expressed us-

demand errors (too many students assigned to a class) and excess supply errors (empty seats in a positively priced class). The Wharton Committee viewed excess demand errors as more costly than excess supply errors and tuned the A-CEEI software accordingly for the experiment. Over the eight sessions, there were 10 total seats of excess supply (median: one seat per session) and two total seats of excess demand (median: zero seats per session). The Pareto-improving trades exercise reported in the text treats the registrar as owning the 10 seats of excess supply and ignores the two seats of excess demand. In the practical implementation of A-CEEI at Wharton, we modified the mechanism in a small way to entirely prevent excess demand errors that cause violations of strict capacity constraints (e.g., due to fire codes). See Budish et al. (2017).

ing the language we provided. If such preferences were present in our subject pool, this would necessarily create a discrepancy between subjects' reported preferences and their true preferences.

Returning to the analogy from the Introduction: if one's true underlying preferences are in English, and the preference reporting language is Latin, the first issue is that translating from English to Latin requires mastery of both English and Latin and skill at translation, whereas the second issue is that there are some concepts and ideas that can be expressed in English that cannot be fully expressed in Latin.

We present summary statistics on use of the preference reporting language and an initial analysis of preference reporting inaccuracies in Section 4.1. We then investigate the two sources of inaccuracies in Sections 4.2 and 4.3. Section 4.4 discusses the results from this section.

4.1 Preference Reporting Language Use and Accuracy

Table 4 presents summary statistics describing how subjects used the preference reporting language and the accuracy of subjects' preference reports.

Panel A of Table 4 shows summary statistics on the use of the preference reporting language. The data suggest that subjects generally followed the instructions we provided. We advised subjects to report positive cardinal values for at least 12 courses. The median number of courses assigned positive values was 12 and a large majority of subjects (76.5%) reported positive values for 11 or more courses. In addition, we advised subjects to assign their favorite course a value of 100 and to assign all other courses a relative value. Again, a large majority of subjects (75.0%) reported a value of 100 for one and only one course. Generally speaking, subjects spread their values of courses evenly from 0 to 100. The last three rows of Panel A suggest that most subjects chose not to use any adjustments (the median subject used zero adjustments) and the average number of adjustments across all subjects was slightly more than one. Of the adjustments that were made, there was an even split between positive adjustments (reflecting complementarity between course sections) and negative adjustments (reflecting substitutability between course sections).

Table 4: Use of the Preference Reporting Language and Contradictions

Panel A: Use of Preference Reporting Language ($n = 132$)

	Mean	Min	25th Pct.	Median	75th Pct.	Max
# courses valued $v > 0$	12.45	7	11	12	14	24
# courses valued $v = 100$	1.40	0	1	1	1	8
# courses valued $50 \leq v \leq 99$	4.87	0	3	5	7	10
# courses valued $0 < v < 50$	6.17	0	4	6	8	17
# adjustments	1.08	0	0	0	2	10
# adjustments > 0 (complements)	0.55	0	0	0	1	10
# adjustments < 0 (substitutes)	0.53	0	0	0	1	6

Panel B: Preference Reporting Contradictions ($n = 1661$, $s = 126$)

	Any contradiction	"Prefer" or "Strongly Prefer"	"Strongly Prefer"
All binary comparisons (1661)	15.59%	10.42%	3.49%
Below 10% utility difference (400)	25.75%	17.00%	6.25%
Above 10% utility difference (1261)	12.37%	8.33%	2.62%
Above 50% utility difference (407)	7.37%	4.67%	1.47%

Notes: Panel A reports on the use of the preference reporting language for the 132 subjects in the experiment. v is the cardinal value assigned to a particular course section. Panel B reports summary statistics on the rate of preference reporting contradictions, defined as a choice from a binary comparison that is inconsistent with the ordinal ranking implied by the subject's preference report. These data cover the 126 subjects for whom we collected binary choice data (see footnote 36 about the other six subjects).

Panel B of Table 4 shows summary statistics about preference reporting accuracy. As discussed above, every binary comparison is a test of our preference reporting language. We say a binary comparison choice is *consistent* if the subject's choice from the binary comparison is what we would expect based on that subject's reported preference data; otherwise, we say it is a *contradiction*. The first observation to make about this data is that there are a significant number of contradictions. Overall, 15.6% of binary comparisons are contradictions (see the first row of Panel B). If we instead look at the data at the subject level rather than the binary comparison level, 75.4% of subjects have at least one contradiction.

That said, and perhaps reassuringly, subjects make relatively few "big" preference reporting mistakes. Whereas 15.6% of binary comparisons are contradictions, just 3.5% of binary comparisons are contradictions in which subjects indicate they Strongly Prefer the schedule their preference reports suggest they like less. Binary comparisons in which the two schedules' cardinal utilities are within 10% of each other (roughly the bottom quartile of utility differences) — suggesting a close call — result in contradictions 25.8% of the time and Strongly Prefer contradictions 6.3% of the time. In contrast, schedules where the utility

difference exceeds 10% result in contradictions just 12.4% of the time and Strongly Prefer contradictions 2.6% of the time. If the utility difference is more than 50% (roughly the top quartile of utility differences), just 7.4% of comparisons are contradictions and just 1.5% are Strongly Prefer contradictions.

In Table 5, we report on the causes of preference reporting inaccuracies using a regression framework. We run Probit regressions with a dependent variable that is equal to 1 if a binary comparison choice is a contradiction and 0 if it is consistent. We regress this dependent variable on various characteristics of the binary comparison in an attempt to understand what causes contradictions. We report marginal effects so that the coefficients can be interpreted as the change in probability of a contradiction and cluster our standard errors at the subject level to account for correlations in the errors for each subject. Appendix G demonstrates the robustness of the results presented in Table 5.

Table 5: Causes of Contradictions

	Dependent Variable: *Contradiction*				
	(1)	(2)	(3)	(4)	(5)
$\lvert log(utility\ A) - log(utility\ B)\rvert$	-0.327	-0.256	-0.329	-0.327	-0.259
	(0.043)***	(0.052)***	(0.043)***	(0.043)***	(0.047)***
Cardinal (369 comparisons)		0.164			0.156
		(0.035)***			(0.035)***
Combinatorial (87 comparisons)			-0.040		0.002
			(0.031)		(0.039)
Lower utility schedule has				0.056	0.044
"elegant" feature (241 comparisons)				(0.035)*	(0.033)
Predicted Probability at Mean Values	0.135	0.137	0.134	0.134	0.129
Observations	1,661	1,574	1,661	1,661	1,661
Clusters (Subjects)	126	122	126	126	126
R-Squared	0.052	0.087	0.053	0.055	0.091

Notes: Probit regressions with a dependent variable equal to 1 if a binary comparison choice was a contradiction and equal to 0 if the binary comparison choice was consistent. Marginal effects are reported. Analyzes the 126 subjects for whom we have binary comparison data and excludes comparisons in which both schedules generate equal cardinal utility (such that the reported preference data does not generate a strict preference between schedules). Standard errors are robust and clustered at the subject level. Significance is denoted with stars: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Column 1 of Table 5 shows a regression of the dummy for a binary comparison choice being a contradiction on the absolute value of the difference between the logarithm of cardinal utility generated by each schedule in the binary comparison. For small utility differences this can be thought of as the percentage change in utility going from the cardinal

utility of the schedule that generates less utility to the cardinal utility of the schedule that generates more utility. As $|log(utility\,A) - log(utility\,B)|$ increases, the likelihood of a comparison being a contradiction drops meaningfully. The interpretation of the coefficient is that each 10 percentage point increase in the utility difference between A and B reduces the likelihood of contradiction by 3.3 percentage points (as compared to the average rate of contradiction of 15.6%). This result echoes the pattern in Table 4 Panel B, which showed that contradictions are much more likely when the utility difference between schedules is small than when the difference is large. Because the utility difference between the schedules is so predictive of the likelihood of a contradiction, we continue to control for it in the other regressions in the table, which explore the two sources of preference reporting inaccuracy and are described below.

4.2 Difficulty Using the Preference Reporting Language

To assess whether agents had difficulty using the preference reporting language we provided, we explore whether they were able to effectively use each of its components: cardinal values to express preferences for individual courses and pairwise adjustments to express certain kinds of complementarities and substitutabilities for pairs of courses. We explore subjects' ability to use each of these components of the language in turn.

To examine subjects' ability to report cardinal item values, we differentiate between the ordinal and cardinal component of a subject's reported preferences for individual courses. For this analysis, we drop the 87 binary comparisons in which one or both schedules triggered an adjustment. For the remaining 1574 binary comparisons, we say that a comparison between schedules A and B is an *ordinal comparison* if the preferences the subject reports imply a preference between schedule A and B based on ordinal information alone. For example, if we rank course sections by a subject's assigned cardinal item values and find that schedule A consists of the subject's $\{1^{st}, 3^{rd}, 5^{th}, 7^{th}, 9^{th}\}$ highest value course sections while schedule B consists of the subject's $\{2^{nd}, 4^{th}, 6^{th}, 8^{th}, 10^{th}\}$ highest value course sections, then we can conclude that the subject prefers schedule A based on ordinal information alone (i.e., we do not need to know the specific cardinal utilities the student assigned to each course). When one schedule can be determined to be preferred to the other based on ordinal information alone, we say that schedule "rank dominates" the other schedule.

We say that a comparison between schedules A and B is a *cardinal comparison* if neither schedule rank dominates the other. For example, if schedule A consists of a subject's $\{1^{st}, 2^{nd}, 8^{th}, 9^{th}, 10^{th}\}$ highest value course sections and schedule B consists of a subject's $\{3^{rd}, 4^{th}, 5^{th}, 6^{th}, 7^{th}\}$ highest value course sections, ordinal information alone is insufficient to determine which is preferred. These are the comparisons for which the subject's ability to

report cardinal preference information accurately is put to the test.

Column 2 of Table 5 shows that cardinal comparisons are more likely to be associated with a contradiction than ordinal comparisons (the excluded group) even controlling for the fact that ordinal comparisons are generally associated with a larger utility difference between the schedules. The interpretation of the coefficient is that a cardinal comparison is 16 percentage points more likely to be a contradiction than an ordinal comparison, which is both economically large, relative to an average rate of contradiction of 15.8% among comparisons of schedules without an adjustment, and highly statistically significant, with a z-stat of 5.52. This result suggests that subjects had meaningful difficulty reporting cardinal utilities.

To examine subjects' ability to report complementarities and substitutabilities, we explore subjects' use of adjustments. Pairwise adjustments were not used as widely as one might have expected — just 1.08 per subject on average as shown in Table 4. Due to the relatively limited use of adjustments, only 87 binary comparisons involved a schedule in which an adjustment was activated. For this analysis, we compare these binary comparisons, which we call *combinatorial comparisons*, to the other 1574 comparisons. Column 3 of Table 5 finds that combinatorial comparisons are directionally, but not significantly, less likely to generate a contradiction. While it is hard to draw conclusions with this data, the result suggests that adjustments did not detract from preference reporting accuracy.

4.3 Limitations of the Preference Reporting Language

The preference reporting language we used in the experiment was not fully expressive (as defined, e.g., in Nisan 2006), meaning that there exist ordinal preferences over schedules that subjects would be mathematically unable to express using the language that was provided. The issue is that many kinds of non-additive preferences cannot be expressed using pairwise adjustments.[39] Additionally, there are many kinds of non-additive preferences that in principle could be expressed using the language but for which the language does not seem especially natural.[40]

[39] We discussed with the Wharton committee whether to allow subjects to express adjustments over arbitrary sets of courses rather than just pairs, which in principle would make the language fully expressive. In these discussions, we and the committee concluded that arbitrary set-wise adjustments would be too complicated for students. How best to trade off the expressiveness of a preference reporting language and agents' ability to use the language is an interesting open area for research, as we discuss further in the Conclusion.

[40] For example, suppose a student wants to express that they want at most one out of a set of k classes. They could express this in principle using just pairwise adjustments, but it would take $k\frac{(k-1)}{2}$ such adjustments (reporting that any two of the k courses together have negative total value). A simpler way to convey the same preferences would be to report a constraint of the form "at most one out of these k", were the ability to do so provided. See Milgrom (2009) for an example of a preference reporting language that allows agents to express preferences of this form — at most k out of set S. There are numerous analogous examples.

The set of potential non-expressible preferences is vast, and we do not have a disciplined way of exploring all such possibilities as a source of preference reporting contradictions.[41] Instead, we explored two specific sources of non-additive preferences that the Wharton committee suggested to us would be the most important, both of which arise from scheduling considerations per se rather than the contents of the classes within the schedule.

The first is whether a student's schedule is *balanced* — at least one class on each day Monday through Thursday (none of the course sections in our experiment met on Friday, as is typical at Wharton). The second is whether the schedule is *contiguous* — every day on which a student has class he has at most one 1.5-hour gap between the start of the first class and the end of that last one. According to the Wharton committee, these characteristics make a schedule "elegant" and are highly valued by at least some students. However, subjects are not able to express a value for either characteristic using the preference reporting language in the experiment. We therefore investigate whether there are more contradictions in binary comparisons in which the schedule that receives a lower utility based on reported preferences is elegant in at least one of these two ways (and so may generate utility that the subject was unable to report using the preference reporting language) and the other schedule is not elegant in that way.

Column 4 of Table 5 shows that comparisons in which the schedule with the lower utility is elegant in a way that the schedule with the higher utility is not are marginally statistically significantly more likely to be a contradiction ($z = 1.75$). That subjects are more likely to make a contradiction when their reported preferences predict they get more utility from a schedule that is not elegant than a schedule that is elegant suggests that at least some of the contradictions are due to the preference reporting language failing to provide a way for agents to report important features of their preferences. An important caveat is that each of these two types of non-expressible preferences (i.e., being balanced and being contiguous) account for only a small number of contradictions each.[42] There are likely many other non-expressible preferences that we do not quantify here.

[41] With roughly 50,000 possible schedules in the lab, there are 50,000! possible ordinal preferences over schedules, or roughly $10^{12,499}$. As such, the up to 19 binary comparisons we ask of subjects do not provide enough data to identify patterns in such a large set without prior guidance on where to look.

[42] There are 15 contradictions in which the lower utility schedule is balanced and the higher utility schedule is not, and 35 contradictions in which the lower utility schedule is contiguous and the higher utility schedule is not. If we run the regression reported in Table 5 Column 4 separately on *balanced* and *contiguous*, the coefficient on *balanced* is 0.15 and significant at the 1% level, and the coefficient on *contiguous* is 0.030 and not significant. The large magnitude on *balanced* suggests this feature may be important to a meaningful proportion of students, but the number of observations is small.

4.4 Discussion

Subjects' preference reports convey significant information about their preferences: 84.4% of binary comparisons are consistent with the preference reports and large mistakes are comparatively rare. At the same time, however, preference reporting difficulties are prevalent: 15.6% of binary comparisons are contradictions and over three-quarters of subjects exhibit at least one contradiction. As shown in Section 3, these difficulties demonstrably harmed mechanism performance.

Results from this section provide some evidence as to the sources of inaccurate preference reporting. First, subjects had particular difficulty with reporting cardinal preference intensity information — controlling for utility differences between schedules, a binary comparison choice was dramatically more likely to be a contradiction when the preference reports relied on cardinal information to determine which schedule was preferred. Second, when subjects expressed non-additive preferences they did so with reasonable accuracy, but they did so rarely, and the evidence suggests that there were some non-additive preferences that were important to subjects but that subjects were unable to express with the language provided.

These results provide empirical support for some common intuitions in the market design literature, such as the ease of reporting ordinal information relative to cardinal information (Bogomolnaia and Moulin 2001), the importance of non-additive preferences (Cantillon and Pesendorfer 2007, Reguant 2014) and the overall importance of language design (Milgrom 2009, 2011). We also hope that the overall logic of the results in this section gives the reader additional comfort as to the validity of the experimental methodology.

It is worth noting that our results on preference reporting also guided practical implementation at Wharton in a few ways. First, Wharton opted to use the same language in practical implementation as was used in the lab, based on the overall level of accuracy of the reports taking into consideration that subjects had only 10 minutes to report their preferences and had only minimal training. Second, Wharton provided students with extensive training on how to use the reporting language with significant training focused specifically on how to think about cardinal preference intensity, since this was such an important source of difficulty in the lab. Third, Wharton enhanced the top-ten widget (cf. Figure 2) in the preference reporting user interface to allow students to see substantially more than 10 schedules, allowing students to assess whether they had reported their preferences accurately not just for their very most preferred schedules (which may be unattainable if the student likes mostly popular courses) but further down their overall ranking as well.[43] To date, Wharton has opted not to incorporate other ways to report non-additive preferences beyond the pairwise adjustment tool, fearing excessive complexity. Developing a conceptual understanding

[43] In the free-response component of our survey, several subjects specifically mentioned the top-ten widget as a helpful feature of the user interface.

of the tradeoff between expressiveness and complexity is an interesting open area for future research.

5 Analysis of Survey Data

As noted in the Introduction, an additional advantage of using real market participants as experimental subjects is that we could search for "side effects" — issues not captured by the theory that could undermine the potential benefits of a new market design. For example, a mechanism might have attractive theoretical fairness properties but market participants might nevertheless subjectively find it to be unfair. A mechanism might have attractive incentive properties but participants might not understand that they should report truthfully (cf. Hassidim, Romm and Shorrer 2016, Li 2017, Rees-Jones 2018, Rees-Jones and Skowronek 2018). Market participants might find a mechanism to be frustrating or confusing, properties that would undermine the practical appeal of a mechanism but that seem difficult to capture in a theory model.

Concern about side effects was especially pronounced in our setting both because of the nature of the mechanism being considered and the nature of the allocation problem. Regarding the mechanism, A-CEEI had never been used before, so lacked reassuring precedent, and it is complex in several ways that intuitively raise concerns about side effects. Regarding the setting, fear that a new market design might lead to unexpectedly dissatisfied market participants was high at Wharton, where student satisfaction is a top priority — the Wharton committee was concerned about student satisfaction both with regard to the final allocation and the process that lead to that allocation.

To address these concerns, we collected a wide variety of survey data to search for issues missed by the theory. After the completion of each individual mechanism, we asked subjects to report their level of agreement, on a seven-point Likert scale from Strongly Disagree to Strongly Agree, with 12 relatively subjective statements such as "The way courses are allocated through this course allocation system is fair" (Q1) and "I like this course allocation system" (Q7). After the completion of both mechanisms we asked subjects three additional questions allowing them to report which system they preferred (Q13), which they thought others would prefer (Q14) and in which system they liked their schedule better (Q15). Last, we provided subjects with the opportunity to provide anonymous free-response comments. The complete list of questions as well as mean responses are provided in Table 6.

Table 6: Survey Responses

Panel A: Surveys After Completion of each Individual Mechanism

Question	A-CEEI mean	Auction mean	Mean difference	p-value
1. The way courses are allocated through this course allocation system is fair	5.21	4.98	0.23	0.19
2. This course allocation system is easy for me to use.	4.79	4.68	0.11	0.60
3. I understand how this course allocation system works.	4.83	5.92	-1.09***	< 0.001
4. This course allocation system led to the best outcome I could hope for.	4.11	4.34	-0.23	0.23
5. I am satisfied with my course outcome.	4.67	5.00	-0.33	0.19
6. I enjoyed participating in this course allocation system.	4.72	4.37	0.35*	0.095
7. I like this course allocation system.	4.55	4.18	0.36*	0.095
8. My fellow students will like this course allocation system.	4.30	4.33	-0.030	0.88
9. I felt like I had control over my schedule in this course allocation system.	3.95	4.45	-0.50*	0.073
10. This course allocation system is simple.	4.45	3.73	0.72***	0.0012
11. I had to think strategically about what other students would do in this course allocation system.	2.93	6.42	-3.48***	< 0.001
12. Someone with perfect knowledge of the historical supply and demand for courses could have had an advantage over me in this system.	3.67	6.04	-2.37***	< 0.001

Panel B: Survey After Completion of Both Mechanisms

Question	Mean	p-value
13. Which course allocation system did you prefer?	4.06	0.77
14. Which course allocation system do you think your fellow students would prefer?	3.80	0.17
15. In which course allocation system did you get a better schedule?	3.90	0.63

Panel C: Free Response

Please use this page to write any additional comments about your experience during this session. These are anonymous comments, so please do not include your name.

Notes: For Panel A, 1 = Strongly Disagree, 7 = Strongly Agree and the mean difference is the mean of the Auction responses subtracted from the mean of the A-CEEI responses. The p-values for the mean difference are from Wilcoxon sign-rank tests on the mean difference variable, testing against a mean difference of zero. For Panel B, 1 = Strongly Prefer the Auction, 7 = Strongly Prefer A-CEEI. Unlike in Panel A, the Wilcoxon signed-rank test in Panel B tests whether the mean is significantly different from 4.00, the midpoint of the Likert response scale. Significance is denoted with stars: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

We organize our discussion of the survey results into three topics: overall satisfaction, strategic simplicity and transparency. In each of the following three subsections, we discuss the questions associated with that topic. Our grouping of questions into topics is relatively

consistent with groupings based on a Principal Components Analysis (PCA) of the survey questions, as described in Appendix H.

Overall Student Satisfaction

Many of the survey questions aimed at assessing students' overall satisfaction with the two mechanisms. These included questions Q1 ("system is fair"), Q4 ("best outcome I could hope for"), Q5 ("satisfied"), Q6 ("enjoyed"), Q7 ("like"), Q8 ("fellow students will like"), Q13 ("prefer"), Q14 ("fellow students would prefer") and Q15 ("better schedule").[44]

Of these, only two questions yielded statistically significant differences and these were only marginally statistically significant. While these two questions (Q6 and Q7) favored A-CEEI, several of the other questions directionally favor the Auction.

Our main takeaway from this set of questions was that there was not some important unmeasured side effect that caused subjects to dramatically prefer either A-CEEI or the Auction that our main efficiency and fairness analyses would have missed. These results also seemed to give comfort to the Wharton committee that there was nothing unexpected about the A-CEEI mechanism that led the Wharton student subjects to dislike the system.

Strategic Simplicity

By far the largest differences between A-CEEI and the Auction concerned two questions about strategic play: "I had to think strategically about what other students would do in this course allocation system" (Q11) and "Someone with perfect knowledge of the historical supply and demand for courses could have had an advantage over me in this system" (Q12). Another question that is perhaps related to strategic simplicity, "This course allocation system is simple" (Q10), also elicited a large difference between the two mechanisms.[45]

These results suggest that subjects broadly understood the claim made in the experimental instructions that the A-CEEI mechanism did not require strategizing. One might be somewhat surprised that the difference between A-CEEI and the Auction on these measures is not even larger. A potential explanation is that at least some of our subjects were reluctant to accept, or did not understand, that the A-CEEI mechanism was not "gameable" like the Auction (cf. Hassidim, Romm and Shorrer 2016, Li 2017, Rees-Jones 2018, Rees-Jones and Skowronek 2018). A lesson for implementation that came out of these survey responses was to do a more thorough job of explaining this fact to students, since understanding

[44] We conducted a principal components analysis (PCA) of the survey data, reported in Appendix H. The first principal component's largest coefficients were all of the questions we interpreted as having to do with overall satisfaction, as well as Q2 and Q9, which have to do with ease of use.

[45] In the PCA referenced in the previous footnote, the largest magnitudes in the second principal component were Q3, Q10, Q11 and Q12, and the largest in the third principal component were Q11 and Q12.

that historical information and strategizing was not necessary for A-CEEI was positively correlated with other measures of satisfaction with A-CEEI.[46]

Transparency

The two questions on the survey on which A-CEEI performed significantly worse than the Auction were "I understand how this course allocation system works" (Q3) and "I felt like I had control over my schedule in this course allocation system" (Q9).

Our interpretation of these results, in conjunction with some of the free-response comments, is that some subjects felt that A-CEEI was a "black box", i.e., non-transparent.[47] The transparency issue constitutes a side effect in that it negatively impacted market participants' evaluation of the mechanism and was not anticipated by the theory.

Wharton acted on this finding in their practical implementation of A-CEEI in two ways. First, Wharton administrators did student-wide presentations about the new mechanism to explain in detail how it works, the theory behind it and the experimental evidence, all in an effort to enhance transparency. Second, Wharton made a simple but important change to the mechanism's user interface. In the user interface implemented in the lab, subjects were shown the schedule they received under A-CEEI but were not shown market-clearing prices. This prevented subjects from understanding why they received their specific schedule and why, for example, they failed to get some particular course they valued highly. In the practical implementation, Wharton modified the user interface so that students are shown the market-clearing prices. Gérard Cachon, the chair of Wharton's Course Allocation Redesign Team, wrote to us in personal correspondence: "I have heard that this makes a difference – some students say 'when I saw the prices, I understood why I got what I got.'"

Gender

At the time of our experiment, the Wharton administration was facing evidence that women at Wharton disproportionately disliked the Auction. A Wharton survey of all second-year students in the year of our experiment found that women reported lower ratings for the effectiveness of the real Wharton Auction than men did (seven-point scale of effectiveness, 4.95 for men vs. 4.28 for women, t-test, two-sided, $p < 0.001$).[48] The administration was therefore interested in whether A-CEEI would also display a gender disparity.

[46] For more discussion of the benefits of strategy-proofness in market design see, e.g., Pathak and Sönmez (2008, 2013), Roth (2008), Azevedo and Budish (Forthcoming) and Li (2017).

[47] In the free responses, one subject wrote: "I like the idea of getting the best schedule I could afford, but didn't like feeling like I wasn't in control. I would feel helpless if I got a schedule that wasn't close to what I preferred." Another wrote: "The course matching system is just a black box where there's one round and we rely on the computer to make judgments for us."

[48] While the survey question asked about "effectiveness" broadly, it was the only question asked about the Auction and so responses are likely to be driven by feelings about the Auction on multiple dimensions.

Our survey questions about the Auction generated the same pattern as the Wharton administration had seen in their data. In response to all 12 of our survey questions about the Auction, the average response for women indicated that they were (at least directionally) less satisfied than men with the Auction. Six of the 12 questions (Q1, Q2, Q3, Q7, Q9, Q11) displayed significant gender differences (t-tests with $p < 0.1$). Compared to men, women statistically significantly liked the Auction less (Q7, $p = 0.021$), thought it was less fair (Q1, $p = 0.087$), found it less easy to use (Q2, $p = 0.065$), understood it less well (Q3, $p < 0.01$), felt less control over the outcome (Q9, $p = 0.057$) and had to think more strategically to use it (Q11, $p = 0.032$). In contrast, there was no systematic gender preference for A-CEEI. Women were at least directionally more satisfied than men with A-CEEI in five of the questions, and men were at least directionally more satisfied than women in seven of the questions. Only two questions displayed statistically significant gender differences, and they went in opposite directions. Women understood A-CEEI less well than men (Q3, $p = 0.070$), but men said they had to think more strategically to use it (Q11, $p = 0.061$).[49]

Eliminating the gender difference that was present in attitudes toward the Auction was a positive side effect of A-CEEI not anticipated by the theory. If we interpret the Auction as "competitive" because it is highly strategic and A-CEEI as "noncompetitive" because it is approximately strategy-proof, the finding echoes a famous finding in the gender literature (Niederle and Vesterlund 2007).

6 Conclusion

Wharton formally decided to adopt A-CEEI for use in practice after a series of administrative meetings in the few months following our experiment. This could not have been an easy decision given the complexity of the A-CEEI mechanism and the lack of direct precedent. Based on our conversations with the committee, our sense is that what ultimately was pivotal in Wharton's decision to adopt A-CEEI was not any one experimental result but rather the full set of experimental results: the efficiency and fairness gains relative to the Auction; the finding that preference reports were on the whole reasonably accurate, with large mistakes comparatively rare; the finding that the efficiency and fairness gains would be meaningfully larger if preference reporting accuracy could be improved; the strategic simplicity gains identified in the survey; and the lack of any unexpected side effects, beyond the transparency issue which the committee felt could be addressed in practice with better

[49]The lack of significant gender differences when evaluating A-CEEI is not about sample size, which is by definition the same for the A-CEEI tests as the Auction tests, nor is it about the p-value cutoff of 0.1. Of the other 10 gender t-tests for A-CEEI, the smallest p-value is 0.276 and six are above 0.5. Meanwhile of the other six gender t-tests for the Auction, three are $p = 0.108$, $p = 0.113$ and $p = 0.171$, and only one is above 0.5. As is obvious from the text above, all of the close p-values involve women being directionally less satisfied than men with the Auction.

communication and some modest changes to the user interface.

Unfortunately, it was not possible to obtain the data that would have been necessary to do a full empirical before-and-after comparison of the two mechanisms.[50] However, the limited data that are available are all consistent with the claims made by the theory and the experiment. One simple way to measure outcome fairness is to look at the distribution of the most popular courses; for any one student we cannot tell if their failure to get popular courses reflects unfairness of the mechanism or their preferences, but the aggregate distribution suggests that A-CEEI improved equity. In the last fall of the Auction, 32% of students got zero of the top 20 most popular courses and 5% got three or more, versus 13% and 0%, respectively, under A-CEEI. That is, under A-CEEI fewer students got none of the most popular courses and fewer (i.e., none) got three or more. Another way to measure outcome fairness is to look at the distribution of the cost of students' final schedules; the Gini index of this distribution went from 0.54 in the last fall of the Auction to 0.32 in the first fall of A-CEEI.[51] In addition, we used school-wide surveys to investigate the change in mechanisms. At our urging, the annual administration survey of the student body added a few questions about course allocation in the last year of the Auction's use, written in such a way that they could be used again in the first year of A-CEEI (which was implemented as "Course Match") with minimal change to language. The percentage of students responding either Agree or Strongly Agree to the statement "I was satisfied with my schedule from {the course auction system / course match}" increased from 45% in 2013 (the last year of the Auction) to 64% in 2014 (the first year of A-CEEI). The percentage responding either Agree or Strongly Agree for the statement "{The course auction, Course match} allows for a fair allocation of classes" increased from 28% to 65%. The percentage of students responding either effective or very effective to the question "Please rate the effectiveness of the {course auction, course match} system" increased from 24% to 53%.

An interesting open question for future research is how to design a better preference reporting language, both in this specific setting and in general. The results of the experiment show that the language used in the lab and adopted for implementation allowed for preference reports that were accurate enough to yield the efficiency and fairness benefits of A-CEEI, but the results do not at all suggest that the language is optimal. One specific

[50] Ideally, we would have used a school-wide survey to obtain true preference from students during the last year of the Auction; this would have allowed us to compare student outcomes from actual play of the Auction to counterfactual play of A-CEEI, analogously to the study conducted by Budish and Cantillon (2012). Unfortunately, the Wharton administration did not want to conduct such a survey, fearing that a survey of students' "true preferences" at the time they were participating in the Auction would have been confusing — especially given that a school-wide announcement had been made concerning the adoption of the new, truthful mechanism. Due to the complexity of the equilibrium of the Auction, it is an open question whether it is possible to infer true preferences from strategic play in the absence of such a survey.

[51] For further details on these data and the engineering details of the practical implementation see Budish et al. (2017).

direction to consider based on the experiment results would be to allow students to report richer kinds of non-additive preferences. A more difficult conceptual question is how to think about the overall tradeoff between a language's expressiveness and its efficacy. Too simple of a language may actually complicate the mechanism for participants, who must struggle with how to translate their real preferences into too simplistic of a language.[52] Too complicated of a language would also be sub-optimal, if participants are unable to effectively "speak" the language. How to design a language that is optimal for a specific setting is a fascinating question in need of a conceptual breakthrough. A perhaps-related question is whether and how to incorporate prior information about the structure of preference heterogeneity in the relevant population into preference reporting. Typically in market design, a mechanism does not assume anything about the agent's preferences that the agent does not explicitly report to the mechanism via the supplied language. Contrast this with, e.g., common practice at e-commerce companies such as Amazon or Netflix, which interact whatever data they gather about any particular user's preferences with their prior on the structure of preferences in the population to form a posterior of that user's type and make recommendations accordingly. That the Wharton committee was able to identify preferences (e.g., about the temporal structure of schedules as described in Section 4.3) that students had difficulty reporting suggests the potential for advancement on this front.

The endowed preferences methodology has been critically important in the history of market-design experiments, tracing all the way to the early double auction experiments of Chamberlin (1948) and Smith (1962), but it was a non-starter for us given the main question our experiment had to answer. We suspect that as market design continues to grow as a field — and as computers become more powerful and decision supports more sophisticated — market designs leveraging complicated preference information will become more common and many other market design researchers will find themselves in our shoes. Our sincere hope for this paper is that other market design researchers can build on the example here and help bring other useful market designs from theory to practice.

[52] A practical example of using a too-simple reporting language is the restriction on the ability of military cadets to trade off years of service against their desired military branch in cadet-branch matching (Sönmez and Switzer 2013). Also related are limitations on the length of preference lists in school choice (cf. Pathak and Sönmez 2013). See also Hatfield and Kominers (2017) who study theoretically how the design of the contract language in many-to-many matching affects whether preferences, as expressed through the language, are guaranteed to be substitutable and to yield a stable match.

References

Abdulkadiroğlu, Atila, Nikhil Agarwal and Parag Pathak. 2017. "The Welfare Effects of Coordinated Assignment: Evidence from the New York City High School Match." *American Economic Review* 107(12):3635–3689.

Abdulkadiroğlu, Atila, Parag Pathak and Alvin E. Roth. 2005. "The New York City High School Match." *American Economic Review: Papers & Proceedings* 95:364–367.

Abdulkadiroğlu, Atila, Parag Pathak, Alvin E. Roth and Tayfun Sönmez. 2005. "The Boston Public School Match." *American Economic Review: Papers & Proceedings* 95:368–371.

Abdulkadiroğlu, Atila, Parag Pathak, Alvin E. Roth and Tayfun Sönmez. 2006. "Changing the Boston School Choice Mechanism: Strategy-proofness as Equal Access." Working Paper.

Abdulkadiroğlu, Atila and Tayfun Sönmez. 2003. "School Choice: A Mechanism Design Approach." *The American Economic Review* 93(3):729–747.

Ausubel, Lawrence M., Peter Cramton and Paul Milgrom. 2006. The Clock-Proxy Auction: A Practical Combinatorial Auction Design. In *Combinatorial Auctions*, ed. Peter Cramton et al. MIT Press pp. 212–259.

Azevedo, Eduardo and Eric Budish. Forthcoming. "Strategy-proofness in the Large." *Review of Economic Studies* .

Bergemann, Dirk and Stephen Morris. 2005. "Robust Mechanism Design." *Econometrica* 73(6):1771–1813.

Bertrand, Marianne and Sendhil Mullainathan. 2001. "Do People Mean What They Say? Implications for Subjective Survey Data." *American Economic Review* 91(2):67–72.

Bogomolnaia, Anna and Hervé Moulin. 2001. "A New Solution to the Random Assignment Problem." *Journal of Economic Theory* 100:295–328.

Budish, Eric. 2011. "The Combinatorial Assignment Problem: Approximate Competitive Equilibrium from Equal Incomes." *Journal of Political Economy* 119(6):1061–1103.

Budish, Eric and Estelle Cantillon. 2012. "The Multi-Unit Assignment Problem: Theory and Evidence from Course Allocation at Harvard." *American Economic Review* 102(5):2237–2271.

Budish, Eric, Gérard Cachon, Judd Kessler and Abraham Othman. 2017. "Course Match: A Large-Scale Implementation of Approximate Competitive Equilibrium from Equal Incomes for Combinatorial Allocation." *Operations Research* 65(2):314–336.

Calsamiglia, Caterina, Guillaume Haeringer and Flip Klijn. 2010. "Constrained School Choice: An Experimental Study." *American Economic Review* 100(4):1860–1874.

Cantillon, Estelle and Martin Pesendorfer. 2007. "Combination Bidding in Multi-Unit Auctions." Working Paper.

Castillo, Marco and Ahrash Dianat. 2016. "Truncation Strategies in Two-Sided Matching Markets: Theory and Experiment." *Games and Economic Behavior* 98:180–196.

Chamberlin, Edward H. 1948. "An Experimental Imperfect Market." *Journal of Political Economy* 56(2):95–108.

Chen, Yan, Ming Jiang, Onur Kesten, Stéphane Robin and Min Zhu. 2018. "Matching in the Large: An Experimental Study." *Games and Economic Behavior* 110:295–317.

Chen, Yan and Tayfun Sönmez. 2006. "School Choice: An Experimental Study." *Journal of Economic Theory* 127(1):202–231.

Cramton, Peter, Yoav Shoham and Richard Steinberg, eds. 2006. *Combinatorial Auctions*. Cambridge, MA: MIT Press.

Ding, Tingting and Andrew Schotter. 2017. "Matching and Chatting: An Experimental Study of the Impact of Network Communication on School-Matching Mechanisms." *Games and Economic Behavior* 103:94–115.

Echenique, Federico, Alistair J. Wilson and Leeat Yariv. 2016. "Clearinghouses for Two-Sided Matching: An Experimental Study." *Quantitative Economics* 7(2):449–482.

Echenique, Federico and Leeat Yariv. 2013. "An Experimental Study of Decentralized Matching." Working Paper.

Ehlers, Lars and Bettina Klaus. 2003. "Coalitional Strategy-Proof and Resource-Monotonic Solutions for Multiple Assignment Problems." *Social Choice and Welfare* 21:265–280.

Featherstone, Clayton and Muriel Niederle. 2016. "Boston versus deferred acceptance n an interim setting: An experimental investigation." *Games and Economic Behavior* 100:353–375.

Foley, Duncan. 1967. "Resource Allocation and the Public Sector." *Yale Economic Essays* 7:45–98.

Fragiadakis, Daniel E. and Peter Troyan. 2016. "Designing Mechanisms to Focalize Welfare-Improving Strategies." Working Paper.

Fudenberg, Drew and Jean Tirole. 1991. *Game Theory*. Cambridge, MA: MIT Press.

Gale, Douglas and Lloyd Shapley. 1962. "College Admissions and the Stability of Marriage." *The American Mathematical Monthly* 69(1):9–15.

Goeree, Jacob K. and Charles A. Holt. 2010. "Hierarchical Package Bidding: A Paper & Pencil Combinatorial Auction." *Games and Economic Behavior* 70:146–169.

Hakimov, Rustamdjan and Onur Kesten. 2018. "The Equitable Top Trading Cycles Mechanism for School Choice." *International Economic Review* 59(4):2219–2258.

Hassidim, Avinatan, Assaf Romm and Ran I. Shorrer. 2016. "'Strategic' Behavior in a Strategy-Proof Environment." Working Paper.

Hatfield, John William. 2009. "Strategy-Proof, Efficient, and Nonbossy Quota Allocations." *Social Choice and Welfare* 33(3):505–515.

Hatfield, John William and Scott Duke Kominers. 2017. "Contract Design and Stability in Many-to-Many Matching." *Games and Economic Behavior* 101:78–97.

Kagel, John H. and Alvin E. Roth. 2000. "The Dynamics of Reorganization in Matching Markets: A Laboratory Experiment Motivated by a Natural Experiment." *Quarterly Journal of Economics* 115(1):201–235.

Kagel, John H., Yuanchuan Lien and Paul Milgrom. 2010. "Ascending Prices and Package Bidding: A Theoretical and Experimental Analysis." *American Economic Journal: Microeconomics* 2(3):160–185.

Kahneman, Daniel, Jack L. Knetsch and Richard H. Thaler. 1990. "Experimental Tests of the Endowment Effect and the Coase Theorem." *Journal of Political Economy* 98(6):1325–1348.

Kapor, Adam, Christopher A. Neilson and Seth D. Zimmerman. 2018. "Heterogeneous Beliefs and School Choice Mechanisms." NBER Working Paper 25096.

Klemperer, Paul. 2004. *Auctions: Theory and Practice*. Princeton and Oxford: Princeton University Press.

Kojima, Fuhito. 2009. "Random Assignment of Multiple Indivisible Objects." *Mathematical Social Sciences* 57(1):134–142.

Krishna, Aradhna and Utku Ünver. 2008. "Improving the Efficiency of Course Bidding at Business Schools: Field and Laboratory Studies." *Management Science* 27(2):262–282.

Levin, Jonathan and Andy Skrzypacz. 2016. "Properties of the Combinatorial Clock Auction." *American Economic Review* 106(9):2528–2551.

Li, Shengwu. 2017. "Obviously Strategy-Proof Mechanisms." *American Economic Review* 107(11):3257–3287.

McKinney, C. Nicholas, Muriel Niederle and Alvin E. Roth. 2005. "The Collapse of a Medical Labor Clearinghouse (and Why Such Failures Are Rare)." *American Economic Review* 95(3):878–889.

Milgrom, Paul. 2004. *Putting Auction Theory to Work*. Cambridge, UK: Cambridge University Press.

Milgrom, Paul. 2009. "Assignment Messages and Exchanges." *American Economic Journal: Microeconomics* 1(2):95–113.

Milgrom, Paul. 2011. "Critical Issues in the Practice of Market Design." *Economic Inquiry* 49:311–320.

Milgrom, Paul and Ilya Segal. Forthcoming. "Clock Auctions and Radio Spectrum Reallocation." *Journal of Political Economy* .

Moulin, Hervé. 1995. *Cooperative Microeconomics*. London: Prentice Hall.

Myerson, Roger. 1991. *Game Theory: Analysis of Conflict*. Cambridge, MA: Harvard University Press.

Narita, Yusuke. 2016. "Match or Mismatch: Learning and Inertia in School Choice." Working Paper.

Niederle, Muriel and Alvin E. Roth. 2009. "Market Culture: How Rules Governing Exploding Offers Affect Market Performance." *American Economic Journal: Microeconomics* 1(2):199–219.

Niederle, Muriel and Lise Vesterlund. 2007. "Do Women Shy away from Competition? Do Men Compete too Much?" *Quarterly Journal of Economics* 122(3):1067–1101.

Nisan, Noam. 2006. Bidding Languages for Combinatorial Auctions. In *Combinatorial Auctions*, ed. Peter Cramton et al. MIT Press pp. 215–232.

Othman, Abraham, Eric Budish and Tuomas Sandholm. 2010. Finding Approximate Competitive Equilibria: Efficient and Fair Course Allocation. In *Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems*.

Pais, Joana and Ágnes Pintér. 2008. "School Choice and Information: An Experimental Study on Matching Mechanisms." *Games and Economic Behavior* 64(1):303–328.

Pápai, Szilvia. 2001. "Strategyproof and Nonbossy Multiple Assignments." *Journal of Public Economic Theory* 3(3):257–271.

Pathak, Parag A. and Tayfun Sönmez. 2008. "Leveling the Playing Field: Sincere and Sophisticated Players in the Boston Mechanism." *American Economic Review* 98(4):1636–1652.

Pathak, Parag A. and Tayfun Sönmez. 2013. "School Admissions Reform in Chicago and England: Comparing Mechanisms by their Vulnerability to Manipulation." *American Economic Review* 103(1):80–106.

Prendergast, Canice. 2017. "The Allocation of Food to Food Banks." Working Paper.

Rassenti, S. J., V. L. Smith and R.L. Bulfin. 1982. "A Combinatorial Auction Mechanism for Airport Time Slot Allocation." *The Bell Journal of Economics* 13(2):402–417.

Rees-Jones, Alex. 2018. "Suboptimal Behavior in Strategy-Proof Mechanisms: Evidence from the Residency Match." *Games and Economic Behavior* 108:317–330.

Rees-Jones, Alex and Samuel Skowronek. 2018. "An Experimental Investigation of Preference Misrepresentation in the Residency Match." *Proceedings of the National Academy of Sciences* 115(45):11471–11476.

Reguant, Mar. 2014. "Complementary Bidding Mechanisms and Startup Costs in Electricity Markets." *Review of Economic Studies* 81(4):1708–1742.

Reny, Philip J. 2017. "Assignment Problems." *Journal of Political Economy* 125(6):1903–1914.

Roth, Alvin E. 2002. "The Economist as Engineer." *Econometrica* 70(4):1341–1378.

Roth, Alvin E. 2008. "What Have We Learned from Market Design?" *The Economic Journal* 118:285–310.

Roth, Alvin E. 2015*a*. *Experiments in Market Design*. Vol. 2 Princeton University Press chapter 5.

Roth, Alvin E. 2015*b*. *Who Gets What—and Why: The New Economics of Matchmaking and Market Design*. Boston: Houghton Mifflin Harcourt.

Roth, Alvin E. and Elliott Peranson. 1999. "The Redesign of the Matching Market for American Physicians: Some Engineering Aspects of Economic Design." *American Economic Review* 89(4):748–782.

Roth, Alvin E., Tayfun Sönmez and Utku Ünver. 2004. "Kidney Exchange." *The Quarterly Journal of Economics* 119(2):457–488.

Roth, Alvin E., Tayfun Sönmez and Utku Ünver. 2005. "Pairwise Kidney Exchange." *Journal of Economic Theory* 125(2):151–188.

Roth, Alvin E., Tayfun Sönmez and Utku Ünver. 2007. "Efficient Kidney Exchange: Coincidence of Wants in Markets with Compatibility-Based Preferences." *The American Economic Review* 97(3):828–851.

Roth, Alvin E., Vesna Prasnikar, Masahiro Okuno-Fujiwara and Shmuel Zamir. 1991. "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study." *American Economic Review* 81(5):1068–1095.

Smith, Vernon. 1962. "An Experimental Study of Competitive Market Behavior." *Journal of Political Economy* 70(2):111–137.

Sönmez, Tayfun and Tobias Switzer. 2013. "Matching with (Branch-of-Choice) Contracts at the United States Military Academy." *Econometrica* 81(2):451–488.

Sönmez, Tayfun and Utku Ünver. 2003. "Course Bidding at Business Schools." Working Paper.

Sönmez, Tayfun and Utku Ünver. 2010. "Course Bidding at Business Schools." *International Economic Review* 51(1):99–123.

ONLINE APPENDIX

A Recruitment Materials

From: Kaufold, Howard
Sent: Thursday, November 17, 2011 3:09 PM
To: whg12; whg13
Subject: Do Wharton Research Study, Get Free Food, and Earn Your Chance at Cash Prize!

Dear Students,

We would like to ask for your help in a research study that is recruiting current Wharton MBA students. The research, conducted by a Wharton faculty member along with one of our curricular committees of faculty, department chairs and students, is attempting to understand the decisions of Wharton MBA students as they relate to pending changes in the MBA program. Through this study we will learn valuable information that we will use to improve the experience of Wharton students for years to come.

We want to emphasize that your participation is strictly voluntary. However, as a token of our appreciation, at the end of each session we will randomly choose two students and each one will receive $250. (Each session will have approximately 20 students.) In addition, we will provide you with lunch (noon sessions) or dinner (6pm sessions). Your help will also be greatly appreciated as we want to ensure that we understand as best as possible the preferences of our MBA students with respect to these important design changes in the MBA program.

The study will last 90 minutes and take place in either Room F80 or F375 of Jon M. Huntsman Hall. Sessions will begin at 12 **noon** and 6**pm** on

Monday 11/21 – F375 JMHH
Monday 11/28 – F80 JMHH
Tuesday 11/29 – F80 JMHH
Wednesday 11/30 – F80 JMHH
Thursday 12/1 – F80 JMHH

Please click http://mktgweb.wharton.upenn.edu/mba-bhlab/ to sign up for any available time slot on one of the days listed above. (You need only participate in one session.)

We understand that this a busy time of the year for all students, but we do very much hope you will be able to help us with this valuable research study for our MBA program. Thanks in advance.

Yours,

Thomas S. Robertson, *Dean* Howard Kaufold, *Vice Dean*

B Subject Representativeness

Subjects were representative of all Wharton MBA students on demographics as well as attitudes towards, and behavior in, the Wharton Auction. Using data provided by the Wharton Dean's Office, Table A1 shows the demographics of our 132 subjects as well as the universe of Wharton MBA students in the 2011-2012 academic year. The final column reports the p-value of either a test of proportions or a t-test comparing our subjects to the universe of students. We see that based on demographics, our subjects are representative of the Wharton student body with $p > 0.1$ for each variable except race.

Important for our purposes, our subjects look identical to the student body with regard to Auction behavior: namely, the number of points they had at the start of the Spring Auction (which began before the study took place) and the number of points they had when our study took place (points in the fourth round of the Spring Auction). For the second-year students in our study, we also examine data on their attitudes towards the Wharton Auction as measured on the preceding spring's stakeholder survey. Our second-year subjects were almost identical to the universe of second-year subjects in reports on the effectiveness of the Wharton Auction.

Table A1: Representativeness of Experimental Subjects

	Subjects (n_s)	Wharton MBAs (n_{mba})	p-value (two-sided)
Panel A: Demographics			
($n_s = 132$, $n_{mba} = 1660$)			
First Year Student	50.8%	51.7%	0.83
Female	47.0%	42.0%	0.26
From United States	37.1%	34.3%	0.52
Finance Major	23.5%	25.7%	0.57
Total Registered Credits	17.1	17.0	0.96
Wharton Credits	11.5	11.3	0.56
White	48.5%	37.2%	0.01***
Asian	20.5%	27.0%	0.10*
Black, Non-Hispanic	5.3%	4.0%	0.46
Hispanic	3.0%	3.4%	0.83
Multi-Race	8.3%	7.2%	0.62
No race reported	14.4%	21.1%	0.07*
GPA	Subjects directionally higher		0.14
Panel B: Auction Behavior			
Points at Start of Spring Auction	6899.6	6966.4	0.79
Points in 4th Round of Spring Auction	4992.3	4960.7	0.92
Panel C: Auction Beliefs			
(Second years only: $n_s = 62$, $n_{mba} = 731$)			
Reported Auction Effectiveness (0 to 7)	4.69	4.68	0.96

Notes: Table A1 reports data provided by Wharton. Due to Wharton's policy of grade non-disclosure, GPA levels cannot be reported. For the variables Total Registered Credits and Wharton Credits in Panel A and the variables in Panel B, the number of MBA students that appear in the mean is 1,649. There are 11 students for whom we have general demographic and GPA information but for whom we do not have auction information. The auction beliefs data in Panel C came from the stakeholder survey completed by rising second year students the preceding spring, so we only have it for the second-year students. Tests are two-sided t-tests (for continuous variables) or two-sided tests of proportions (for binary variables).

C Study Instructions

Study Instructions

Thank you for participating in this study.

If you have a question about the study at any time, please raise your hand.

In this study you will be constructing hypothetical class schedules for the spring semester of your second year at Wharton.

You will construct a schedule twice, once under each of two different course allocation systems.

One course allocation system is a simplified version of Wharton's current MBA "Course Auction". The other is an alternative course allocation system for Wharton MBA courses called the "Course Matching System".

Half the sessions in the study will use the "Course Auction" first and half will use the "Course Matching System" first.

After you construct a schedule under each system, you will answer a series of questions about the schedule you have constructed and about the system that you used.

After you have constructed schedules under both systems, you will be asked to compare around 15 to 20 pairs of schedules. For each pair of schedules you will be asked which of the two you prefer.

While using each system, please imagine that it is the spring term of your second year at Wharton, so this will be your last chance to take Wharton classes. Please try to construct your most preferred schedule given the courses that are available.

We are using a subset of 25 spring semester course sections. These course sections were selected to be representative in terms of scheduling, department, and popularity level.

There may be some courses that you would be interested in taking that are not included on this list. There is a limited set of courses because there are only approximately 18 students in the study today and so we cannot replicate the entire course offerings of a normal spring semester. (Note that the actual roster for this spring may differ in terms of which courses are offered, the professors teaching them, and their meeting times.)

We ask you to imagine that these are the only courses available in the spring semester of your second year at Wharton, and to construct your most preferred schedule given these courses. Since this is your last semester, any budget points that you do not use are worthless.

Please imagine that you do not need to take any particular courses for your major or any other graduation requirements, but that you do need to take 5 credit units. If you have already taken one of the courses in the sample, then you should assume that you cannot take the course again in the spring semester. On the other hand, you should assume that you can take any course in the sample that you have not already taken, that is, ignore any prerequisite requirements. Notice that all of the courses are semester length and worth one credit unit.

Imagine that this is the schedule you would construct the week before classes begin. Once classes start you would be able to drop a course, but you would have to replace it with a course that had an open seat.

In real life, we know you take these decisions very seriously. We ask that you take the decisions in this session seriously as well. We will provide you with time to think carefully while using each system.

Note: Neither the schedules you construct nor the decisions you make in this experiment will have any impact on your actual spring semester courses or your point budget in the actual Wharton MBA Course Auction.

The course sections that are available are listed in the packet that has been given to you. Please take five minutes to look through the packet of courses that are available. Think about how interested you are in each of the courses and what would be your ideal schedule or schedules. We will begin with the first system in five minutes.

Instructions for the Course Auction

This procedure is a simplified version of Wharton's current MBA Course Auction. It is similar to the Course Auction that you have already used during your time at Wharton, but with a few differences:

- Every student starts with the same number of budget points (5,000)
- There are 4 rounds of auction activity
- All students are considered second-year students bidding on courses for their last semester
- All students need 5 credit units (CUs)

You are given a budget of 5,000 points. There are then 4 rounds of the auction, all of which we will play today. In the first round you can bid on as many courses as you would like so long as the sum of your bids is less than or equal to your budget. In the next three rounds, you can buy and sell courses with other students.

Instructions for Round 1

Submitting Bids

In the first round, you can submit bids for as many different course sections as you like. The sum of your bids cannot exceed your budget of 5,000 points.

How are prices calculated?

Prices are calculated the same way as in the current Wharton Course Auction. The price of a section is set at the highest losing bid or 100 points, whichever is higher. For example, if a section has 5 seats, the price for the section is set equal to the sixth highest bid for it, if that bid is at least 100 points, otherwise the price is 100. For example, if the sixth highest bid is 120, then the five highest bidders would each get a seat and be charged 120 points. If fewer students bid for a section than it has seats, then the price of the section is set to 100.

What sections do I get?

You get any section for which your bid is greater than or equal to the price. In the event of a tie, where two or more students submit exactly the same bid and there is not enough space for all of them, the computer randomly assigns the available seats to students who bid that amount.

What happens to my budget?

For each section that you receive, your budget will be decreased by the price of the section. For example, if you bid 1000 for the only section of Course A and its price is 400, then you will receive a seat in Course A, and your budget will be decreased by 400 points. If you do not get a seat in the course then you will not give up those 400 points.

Instructions for Rounds 2, 3, and 4

Submitting Bids and Asks

In Rounds 2 through 4, you can submit bids for as many different sections as you like, just as in Round 1. You can also submit asks, which are offers to sell, for any section that you currently have. The sum of your bids cannot exceed your current budget. You can ask whatever amount you like.

How are prices calculated?

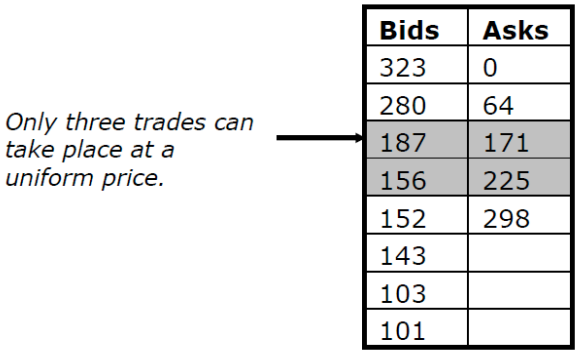
For any section where there are both bids and asks, a trading price is set if there is at least one bid higher than the lowest ask. When this is the case, the computer sets a price to make as many trades as possible. This involves finding a price such that the number of bids higher than that price is the same as the number of asks lower than that price.

Suppose the following bids and asks are submitted for a section during a round.

Bids: 101, 323, 143, 103, 187, 280, 156, and 152.

Asks: 225, 64, 298, 171, and 0.

To see which bids and asks are successful and what the clearing price is, first arrange all the bids in descending order and the asks in ascending order as shown in the table below:



Since only the top three bids are higher than the three lowest asks (and the fourth highest bid is lower than the fourth lowest ask), only three trades can go through. The clearing price is determined as the larger of the first losing bid and the highest winning ask; in this case, the first losing bid is 156, and highest winning ask is 171 — hence the clearing price is 171. The clearing price amount is transferred from each of the successful bidders to each successful seller (the accounts of unsuccessful bidders and sellers remain unaffected).

If there are extra seats in a section, for example if a section does not reach capacity in Round 1, then those seats are treated as if they are being offered for an ask of 100 points.

You can always be guaranteed to drop a section by submitting an ask of "0".

What should my schedule look like at the end of Round 4?

At the end of Round 4 you should have: (1) no more than 5 credit units in your schedule; (2) no sections that have a time conflict with each other; and (3) no more than one section in each course.

Is my schedule after Round 4 my final schedule?

Not necessarily. Recall, you should imagine that this is the schedule you would construct the week before classes begin. Once classes start you would be able to drop a course, but you would have to replace it with a course that had an open seat.

If you have any questions, please raise your hand.

Instructions for Between Systems

You have just constructed a schedule under the first system and answered some questions about the schedule and the system. You will now construct a schedule under the other system.

You are constructing a schedule in this system starting "from scratch" such that the decisions you and the other students in this session made while using the first system do not affect anything about activity in this system.

You should again construct the best schedule you can for your spring term of your second year at Wharton. The same course sections are available for this system as were available for the last one.

Instructions for the Course Matching System

The Course Matching System is different from the Wharton Course Auction with which you may be familiar.

The Course Matching System works differently from an auction in that you do not directly bid for course sections. Instead, the computer acts as your agent to buy the best schedule of courses you can afford.

Your job is to tell the computer how much you value individual course sections and whether you assign extra value (or negative value) to having certain course sections together. This process will be explained in detail below.

Since you can tell the computer how much you like every course or pair of courses that might be in your schedule, the Course Matching System only needs one round. In that round, the computer will use your preferences to buy you the best schedule you can afford.

Since the computer is going to optimally buy courses for you, your job is to provide the computer with all the information it needs about how much you value the courses. This is obviously very important, since the computer is going to buy the optimal schedule for you given only what it knows about how you value courses.

The way to communicate your values to the computer is as follows:

1) **You tell the computer how much you value each course section that you have any interest in taking.**
   - First, you pick a favorite course section and assign it a value of 100.
   - Second, you assign all other course sections that you have any interest in taking a value between 1 and 100.

The reason that you assign your favorite course section a value of 100 and all other sections a number between 1 and 100 is that all values are relative.

For example, if you value every course at 100 then you are telling the computer that you value all courses equally. If you value one course at 100 and another course at 50, you are telling the computer you value the course at 100 twice as much as the course at 50.

Unlike using other course allocation systems, when using the Course Matching System, you do not need to think about what other people are doing. All you need to do is communicate how you value course sections to the computer so it knows how to make tradeoffs for you.

How does assigning value to courses work?

Suppose that among the many course sections you assign a positive value, you tell the computer the following values for the single section courses A through E:

Course A = 100
Course B = 80
Course C = 60
Course D = 15
Course E = 10

This tells the computer that you are particularly interested in Courses A, B and C, and somewhat interested in Courses D and E. In particular, it tells the computer that you prefer getting Courses A, B, and C (100 + 80 + 60 = 240) than getting Courses A, D, and E (100 + 15 + 10 = 125).

It also tells the computer that you prefer getting Courses B and C (80 + 60 = 140) than Courses A, D, and E, which only sum to 125. For any two schedules, the computer thinks you prefer whichever schedule has a larger sum.

For simplicity, this example valued only 5 course sections. You should list a positive value for as many courses that you have any interest in taking. We recommend that you assign a positive value to at least 12 course sections. This way the computer can distinguish between a section that has low positive value to you and a section that has zero value to you.

Can I assign values for multiple sections of the same course?

Yes, and you will probably want to do this. To explain, suppose three sections of a course are offered, all on Mondays and Wednesdays. Professor Smith teaches the 10:30-12:00 and 12:00-1:30 sections while Professor Jones teaches the 3:00-4:30 section. You may assign values of 90, 80 and 15 to these three sections, respectively, to signify that you greatly prefer Professor Smith to Professor Jones, and slightly prefer 10:30 to 12:00. Because you can only take one section of a course, you will be assigned at most one of these three course sections, even though you entered values for all three.

Again, there is no limit to the number of course sections that you may assign a positive value.

2) You tell the computer if you assign extra (or negative) value to certain pairs of classes.

To do this, you check the boxes next to any two sections and indicate an extra positive or negative value to having both sections together. These "adjustments" are shown at the top of the page of your valuations.

Why might I assign extra value to two courses together?

Some students might get extra value from having two courses that are back-to-back in their schedule (e.g. they do not like breaks between classes).

Some students might get extra value from having two courses that are related in their schedule (e.g. they might get extra value from taking two courses from the same department if each one becomes more useful with the other).

You can think of these courses as complements, i.e. the combination of the two courses together is greater in value than the sum of their values.

How does assigning extra value work?

Suppose you specify the following values for single section courses A through C:

Course A = 40
Course B = 30
Course C = 85

And suppose you assign an extra value of 20 for getting Course A and Course B together.

Then you are telling the computer that getting Course A and Course B together in your schedule has a value of 90 (90 = 40 for Course A + 30 for Course B + 20 for getting both together).

This means that the computer would try to get you Course A and Course B together before trying to get you Course C. If you had not assigned the extra value to Courses A and B together, the computer would have tried to get you Course C before trying to get you Courses A and B.

Why might I assign negative value to two courses together?

Some students might get negative value from having two courses that are back-to-back in their schedule (e.g. they prefer to take breaks between classes).

Some students might get negative value from having two courses that are related in their schedule (e.g. they might decide that they only want to take one class from a certain department).

You can think of these courses as substitutes, i.e. the second course is worth less when you already have the first.

How does assigning negative value work?

Suppose you specify the following values for single section courses A through C:

Course A = 40
Course B = 30
Course C = 55

And suppose you assign a negative value of -20 for getting Course A and Course B together.

Then you are telling the computer that getting Course A and Course B together in your schedule has a value of 50 (50 = 40 for Course A + 30 for Course B - 20 for getting both together).

This means that the computer would try to get you Course C before getting you Course A and B together. If you had not assigned the negative value to Courses A and B together, the computer would have tried to get you Courses A and B before trying to get you Course C.

You can also use an adjustment to tell the computer "I want to take at most one of these two courses". Using the example above, suppose you want to take either Course A or Course B, but you absolutely do not want to take both. Then you should assign a negative value of -70 for Course A and B together. That negative adjustment tells the computer that the combination has value 0 to you (0 = 40 for Course A + 30 for Course B – 70 for getting both together). Therefore, you may get Course A or Course B, but the computer will never get both for you.

When do I not need to enter in an adjustment?

You do not need to enter an adjustment when two sections are from the same course or two sections are offered at the same time. The computer already knows that you cannot take these sections together. For example, if Professor Baker teaches two sections of the same course, one from 9:00-10:30 and the other from 10:30-12:00, then you can assign a positive value for each of them, but you don't need to assign a positive or negative adjustment for the combination.

Once the computer knows how much you value each course section, it will buy the best schedule you can afford.

How do I know that I am reporting my values right?

To help make sure you are reporting your values right, you can click a button on the navigation bar to see your top 10 schedules. Given the values you reported, the computer thinks that these are your 10 favorite schedules, ranked in order. This means that the computer will try to buy you these schedules in this order. If the order of these schedules does not look right to you, go back and adjust your values until they appear in the right order.

What is my budget that the computer will use to buy courses for me?

Each student is given a budget of 5,000 points.

How are prices determined?

The Course Matching System sets prices based on demand for the courses so that demand equals supply. Courses that are more highly demanded get higher prices and courses that are less popular get lower prices or prices of zero.

One way to think about how prices are set is that each student's computer asks for the best possible schedule for its student. When everyone has their best possible schedule, some courses will have too many students. The price of those courses will rise. Then, given the new set of prices, each student's computer asks again for the best possible schedule for its student at the new set of prices. Some courses will be undersubscribed or oversubscribed and prices will adjust again. This process repeats until there is a set of prices where all popular courses are full and every student gets their best possible schedule given those prices.

Given the set of prices, it may be necessary to break a tie between two or more students who want a course section. These potential ties are broken by assigning a randomly selected small budget increase to each student.

Shouldn't the values I report to the computer depend on the prices of courses or other student's values?

No! The Course Matching System is designed so you do not need to think about the prices of the courses or the values that other students assign to courses. You get the best schedule possible simply by telling the computer your true values for courses.

To see this, notice that if your favorite course, to which you assign a value of 100, is a course whose demand is less than the number of available seats, then it will have a price of zero and you will get that course without using any of your budget. The computer can then use the remainder of your budget to try to get the other course sections that you value highly.

Another way to think about reporting your values to the computer is to imagine you are sending the computer to the supermarket with your food budget and a list of your preferences for ingredients for dinner. You want to report your true values so that the computer can make the right tradeoffs for you when it gets to the supermarket and observes the actual prices for each ingredient.

Are my values equivalent to "bids"?

No! As mentioned above your values are only compared to each other and never compared with other students' values.

Is the schedule I receive after I report my values my final schedule?

Not necessarily. Recall, you should imagine that this is the schedule you would construct the week before classes begin. Once classes start you would be able to drop a course, but you would have to replace it with a course that had an open seat.

If you have any questions, please raise your hand.

Please use this page to write any additional comments about your experience during this session. These are anonymous comments, so please do not include your name.

D List of Course Sections Available in Experiment and Excerpt of Course Descriptions

At the beginning of each session, along with the instructions reproduced as Appendix C, we distributed to students the list of course sections available in the experiment as well as course descriptions. This list and the first two course descriptions are reproduced below and on the following page. The number of available seats was selected by the Wharton Committee to create scarcity in the laboratory environment anticipating 20 subjects per session. Our actual turnout varied between 14-19 subjects per session. In order to maintain scarcity with fewer subjects we adjusted course capacities as follows. If 18-19 subjects attended, we used the capacities below (107 seats total). If 16-17 subjects attended, we turned five-seat courses into four-seat courses (97 seats total). If 14-15 subjects attended we turned five-seat courses into four-seat courses and turned four-seat courses into three-seat courses (86 seats total).

Course	Title	Instructor	Day Code	Start Time	Stop Time	Available Seats
ACCT742	PROBLEMS IN FIN REPORTIN	LAMBERT R	MW	0130PM	0300PM	5
ACCT897	TAXES AND BUS STRATEGY	BLOUIN J	MW	1200PM	0130PM	4
FNCE726	ADVANCED CORP FINANCE	VAN WESEP,E	TR	1200PM	0130PM	5
FNCE728	CORPORATE VALUATION	CICHELLO M	MW	0300PM	0430PM	4
FNCE750	VENT CAP & FNCE INNOVAT	WESSELS D	MW	0130PM	0300PM	4
FNCE750	VENT CAP & FNCE INNOVAT	WESSELS D	MW	0300PM	0430PM	4
FNCE891	CORPORATE RESTRUCTURING	JENKINS M	TR	0130PM	0300PM	4
LGST806	NEGOTIATIONS	DIAMOND S	R	0300PM	0600PM	3
LGST806	NEGOTIATIONS	BRANDT A	W	0300PM	0600PM	3
LGST809	SPORTS BUSINESS MGMT	ROSNER S	TR	0300PM	0430PM	5
LGST813	LEG ASP ENTREPRENRSHP	BORGHESE R	M	0300PM	0600PM	5
MGMT691	NEGOTIATIONS	MUELLER J	TR	1030AM	1200PM	3
MGMT721	CORP DEV: MERG & ACQUIS	CHAUDHURI S	TR	0900AM	1030AM	4
MGMT721	CORP DEV: MERG & ACQUIS	CHAUDHURI S	TR	1030AM	1200PM	4
MGMT782	STRATEGIC IMPLEMENTATION	MURMANN J	TR	1200PM	0130PM	5
MGMT833	STRAT & PRAC OF FAMILY	ALEXANDER W	TR	0130PM	0300PM	4
MKTG756	MARKETING RESEARCH	IYENGAR R	MW	1030AM	1200PM	5
MKTG773	CUSTOMER BEHAVIOR	REED A	TR	1030AM	1200PM	5
MKTG776	APPL PROB MODELS MKTG	FADER P	W	0300PM	0600PM	5
MKTG778	STRATEGIC BRAND MGMT	MOGILNER C	TR	0130PM	0300PM	5
OPIM690	MANAG DECSN MAKING	MILKMAN K	MW	0130PM	0300PM	5
OPIM692	ADV TOPICS NEGOTIATION	SCHWEITZER M	TR	0130PM	0300PM	4
REAL721	REAL ESTATE INVESTMENTS	FERREIRA F	MW	0130PM	0300PM	4
REAL721	REAL ESTATE INVESTMENTS	WONG M	TR	0130PM	0300PM	4
REAL821	REAL ESTATE DEVELOPMENT	NAKAHARA A	W	0300PM	0600PM	4

ACCT742: PROBLEMS IN FIN REPORTIN - LAMBERT R

Financial statements are a primary means for firms to communicate information about their performance and strategy to investors and other groups. In the wake of numerous accounting scandals and the recent financial meltdown (which accounting both helped and hindered), it is more important than ever for managers and investors to understand (i) the financial reporting process, (ii) what financial statements do and do not contain, and (iii) the types of discretion managers have in presenting transactions they have undertaken. This course is designed to help you become a more informed user of accounting numbers by increasing your ability to extract, interpret, and analyze information in financial statements.

While this is not a course in equity valuation per se, equity valuation is one of the most common uses of financial statement data. Accordingly, we will examine the relation between Accounting 742 - stock prices and financial statement information. We will also study the use of financial ratios and forecasted financial statement data in models of distress prediction.

ACCT897: TAXES AND BUS STRATEGY - BLOUIN J

Traditional finance and strategy courses do not consider the role of taxes. Similarly, traditional tax courses often ignore the richness of the decision context in which tax factors operate. The objective of this course is to develop a framework for understanding how taxes affect business decisions.

Part of being financially literate is a having a basic understanding of how taxation affects business decisions that companies typically face: forming the business and raising capital, operating the firm, distributing cash to shareholders through dividends and share repurchases, expanding through acquisition, divesting lines of business, and expanding internationally. Taxes have a direct impact on cash flow and often divert 40% to 50% of the firm’s pretax cash flow to the government. Having an understanding of taxation and how firms plan accordingly is important whether you will be running the firm (e.g., executive in large company, entrepreneur, or running a family owned business) or assessing it from the outside (e.g., financial analyst, venture capitalist, or investment banker). Taxes are everywhere and it pays to have some understanding of them.

E Order Effects of Main Results

In four of our eight sessions, subjects used the Auction first; in the other four sessions, subjects used A-CEEI first. If using A-CEEI forces subjects to think about their preferences more deeply than using the Auction — and this deeper thought contributes to better outcomes — then we might expect A-CEEI to do particularly well relative to the Auction when subjects use the Auction before A-CEEI (i.e., before they have engaged in the deep thought) as compared to when they use the Auction after A-CEEI. In Table A2 we replicate Table 1 for the two orders and show that our main efficiency and fairness results are quite similar, regardless of which mechanism was used first. If anything, results are directionally stronger when A-CEEI is played first, although differences are far from significant.

Table A2: Binary Comparison and Mechanism Ordering

		Aggregation Level				
		Individual-Subject			*Market-Session*	
			A-CEEI First	*Auction First*	*A-CEEI First*	*Auction First*
Outcome	**Data**					
Efficiency	*Binary Comparison*	(A)			(B)	
		Prefer A-CEEI	27	29	Prefer A-CEEI	4	2
		Prefer Auction	20	22	Prefer Auction	0	0
		Identical outcomes	9	8	Tie	0	2
		Indeterminate preference	10	7			
		p-value	$p = 0.191$	$p = 0.201$	p-value	$p = 0.063$	$p = 0.250$
	Reported Preference	(C)			(D)		
		Prefer A-CEEI	41	38	Prefer A-CEEI	3	4
		Prefer Auction	16	19	Prefer Auction	0	0
		Identical outcomes	9	8	Tie	1	0
		Indeterminate preference	0	1			
		p-value	$p < 0.001$	$p = 0.008$	p-value	$p = 0.125$	$p = 0.063$
Fairness	*Binary Comparison*	(E)			(F)		
		Less Envy A-CEEI	22	18	Less Envy A-CEEI	3	2
		Less Envy Auction	11	12	Less Envy Auction	1	0
		No Envy either	31	34	Tie	0	2
		Same Envy both	2	2			
		p-value	$p = 0.040$	$p = 0.181$	p-value	$p = 0.313$	$p = 0.250$
	Reported Preference	(G)			(H)		
		Less Envy A-CEEI	14	21	Less Envy A-CEEI	4	4
		Less Envy Auction	1	3	Less Envy Auction	0	0
		No Envy either	51	42	Tie	0	0
		Same Envy both	0	0			
		p-value	$p < 0.001$	$p < 0.001$	p-value	$p = 0.063$	$p = 0.063$

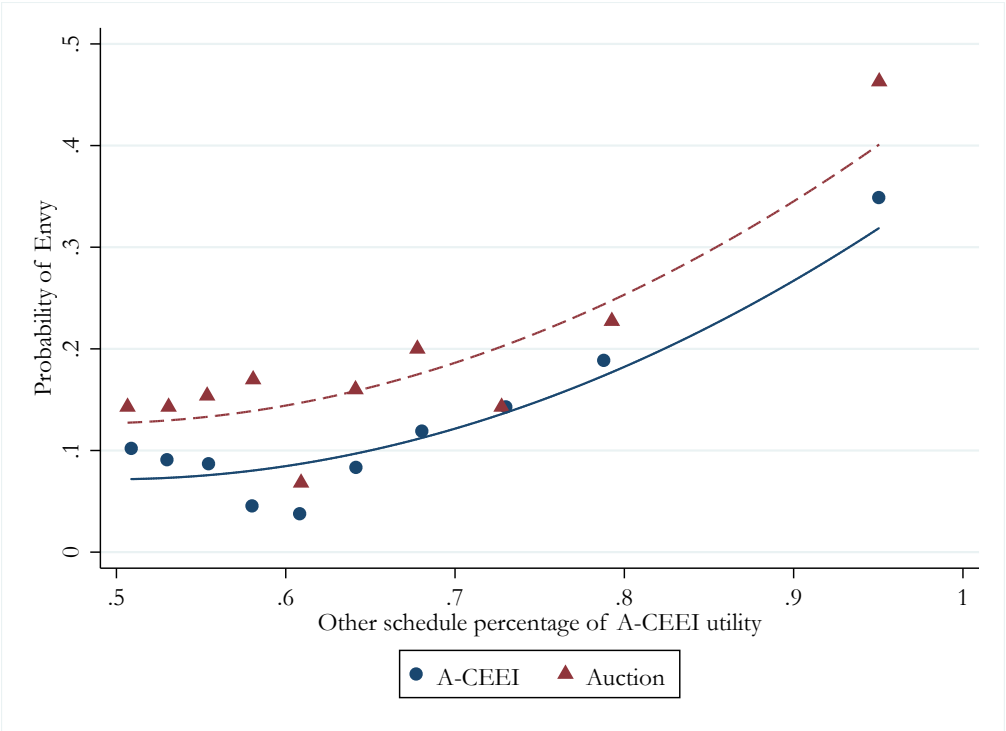
Notes: See notes to Table 1 in the main text. *A-CEEI First* indicates data comes from the four sessions in which subjects used A-CEEI before the Auction. *Auction First* indicates data comes from the four sessions in which subjects used the Auction before A-CEEI.

F Assessing Our Envy Assumption

As discussed in the main text, our envy comparisons involved asking subjects to compare their realized schedule from a mechanism to schedules received by other subjects in the session from the same mechanism. To increase the chance of detecting envy, we selected schedules from the set of others' schedules that delivered at least 50% of the cardinal utility of the subject's A-CEEI schedule (i.e., based on the subject's reported preferences). This design choice allowed us to ensure that we were showing subjects relevant schedules, but it made an implicit assumption that schedules with less than 50% of the cardinal utility of the subject's A-CEEI schedule would not be envied. Here, we assess that assumption.

Figure A1 shows a binned scatter plot of the envy comparisons faced by subjects in the experiment. The graph shows the probability of a subject displaying envy and how it varies with the percentage of the subject's A-CEEI schedule cardinal utility generated by the other subject's schedule in the envy comparison.

Figure A1: Envy as a function of other schedule's utility



Notes: Envy binary comparison data from each mechanism is split into 10 bins based on the percentage of A-CEEI utility generated by the other subject's schedule. The x-axis value is the mean percentage of A-CEEI schedule utility in the bin (to prevent outliers from affecting the location of the highest utility bin, "Other schedule percentage of A-CEEI utility" is Winzorized to the 99^{th} percentile value of 124% of A-CEEI utility). The y-axis value reflects the percentage of envy comparisons in the bin in which the subject at least weakly preferred another subject's schedule to their own. A separate quadratic fit is shown for each mechanism.

Three results are apparent from Figure A1. First, as the utility from the other schedule decreases (as a percentage of a subject's A-CEEI schedule utility), the likelihood that the subject experiences envy falls. In the bin with the 10% highest other schedule utility on the far right of the figure, envy occurs roughly 40% of the time. As the utility from the other schedule decreases (moving left along the figure), envy rates fall in both mechanisms. Second, the data suggest a leveling off of envy rates near 10% (i.e., slightly below 10% for A-CEEI and slightly above 10% for the Auction). This result suggests that as the utility of the other subject's schedule falls to 50% of the subject's A-CEEI utility, the rate of envy does not go to zero (i.e., it remains positive). Third, the envy rate remains higher in the Auction than in A-CEEI, even as the other schedule's utility decreases to 50% of the subject's A-CEEI utility.

These three results allow us to assess our assumption and how it may affect our envy estimates. The first result, that envy is decreasing in the utility of the other schedule, is consistent with the underlying intuition of the assumption. The second result, however, suggests that the assumption does not strictly hold — we would likely have observed at least some envy if we had shown subjects binary comparisons that included schedules that delivered less than 50% of the subject's A-CEEI utility. Nevertheless, the third result suggests that our assumption likely works against us finding that A-CEEI generates less envy than the Auction, since showing subjects additional schedules with lower utility would likely generate higher rates of envy in the Auction than in A-CEEI.

G Robustness of Contradiction Analysis

To show the robustness of our contradiction analysis we replicate the results on the causes of contradictions presented in Table 5 of Section 4, providing three types of robustness tests. We include one table for each set of tests.

First, Table A3 shows the same specifications as Table 5 but also includes a dummy for each subject, to control for potential differences in preference reporting ability across subjects. Given the Probit specification, this analysis effectively drops subjects who never experience a contradiction, narrowing our focus to the 1336 binary comparisons made by subjects who have at least one contradiction. We again report marginal effects so that the coefficients can be interpreted as the change in probability of a contradiction, and we cluster at the subject level. Compared to Table 5, coefficients do not change much and significance levels do not change at all.

Table A3: Robustness of Causes of Contradictions – Subject Dummies

	Dependent Variable: *Contradiction*				
	(1)	(2)	(3)	(4)	(5)
$\lvert log(utility\ A) - log(utility\ B)\rvert$	-0.416	-0.336	-0.415	-0.421	-0.336
	(0.052)***	(0.068)***	(0.052)***	(0.051)***	(0.058)***
Cardinal (369 comparisons)		0.187			0.182
		(0.045)***			(0.045)***
Combinatorial (87 comparisons)			-0.038		0.012
			(0.069)		(0.078)
Lower utility schedule has				0.085	0.081
"elegant" feature (241 comparisons)				(0.051)*	(0.052)
Predicted Probability at Mean Values	0.141	0.148	0.141	0.139	0.135
Subject Dummies	Yes	Yes	Yes	Yes	Yes
Observations	1,336	1,258	1,336	1,336	1,336
Clusters (Subjects)	95	92	95	95	95
R-Squared	0.147	0.181	0.148	0.152	0.183

Notes: See notes to Table 5. Regressions include dummies for each subject.

Second, Table A4 shows the same specifications as Table 5 but includes additional controls. In particular, it includes a dummy variable for the order in which the binary comparison appeared, a dummy for the type of binary comparison (i.e., whether it was an envy comparison, etc.) and a dummy for each session. We again report marginal effects so that the coefficients can be interpreted as the change in probability of a contradiction and cluster at the subject level. Again, compared to Table 5, coefficients do not change much

and significance levels do not change at all.

Table A4: Robustness of Causes of Contradictions – Additional Controls

	Dependent Variable: *Contradiction*				
	(1)	(2)	(3)	(4)	(5)
$\lvert log(utility\ A) - log(utility\ B)\rvert$	-0.380	-0.274	-0.379	-0.377	-0.277
	(0.052)***	(0.060)***	(0.052)***	(0.052)***	(0.055)***
Cardinal (369 comparisons)		0.166			0.156
		(0.033)***			(0.033)***
Combinatorial (87 comparisons)			-0.032		0.008
			(0.030)		(0.036)
Lower utility schedule has				0.053	0.040
"elegant" feature (241 comparisons)				(0.033)*	(0.031)
Predicted Probability at Mean Values	0.127	0.131	0.127	0.127	0.123
Additional Controls	Yes	Yes	Yes	Yes	Yes
Observations	1,661	1,574	1,661	1,661	1,661
Clusters (Subjects)	126	122	126	126	126
R-Squared	0.072	0.107	0.073	0.076	0.110

Notes: See notes to Table 5. Additional controls include dummies for question number (i.e., the order among the binary comparison questions), the type of binary comparison (i.e., an envy comparison, a budget comparison, etc.), and dummies for session.

Third, Table A5 shows the same specifications as Table 5 but only treats comparisons as contradictions if they have a response of Prefer or Strongly Prefer. Results are similar.

Table A5: Robustness of Causes of Contradictions – Prefer or Strongly Prefer

	Dependent Variable: *Contradiction*				
	(1)	(2)	(3)	(4)	(5)
$\lvert log(utility\ A) - log(utility\ B)\rvert$	-0.226	-0.169	-0.227	-0.226	-0.172
	(0.033)***	(0.038)***	(0.033)***	(0.033)***	(0.035)***
Cardinal (369 comparisons)		0.120			0.114
		(0.030)***			(0.030)***
Combinatorial (87 comparisons)			-0.012		0.023
			(0.030)		(0.037)
Lower utility schedule has				0.029	0.021
"elegant" feature (241 comparisons)				(0.029)	(0.026)
Predicted Probability at Mean Values	0.088	0.088	0.088	0.088	0.083
Observations	1,661	1,574	1,661	1,661	1,661
Clusters (Subjects)	126	122	126	126	126
R-Squared	0.047	0.081	0.047	0.049	0.084

Notes: See notes to Table 5. To be a contradiction in this specification requires a response of Prefer or Strongly Prefer.

H Principal Components Analysis of Survey Data

The following table reports a Principal Components Analysis (PCA) of the survey data reported in the main text as Table 6. The PCA illustrates which survey questions tend to move together in subjects' responses. For questions 1-12 the PCA utilizes the difference between the subject's A-CEEI response and Auction response.

Table A6: Principal Components Analysis of Survey Data

	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8
12 Difference Variables								
1. The way courses are allocated through this course allocation system is fair.	**0.258**	-0.190	0.082	-0.111	0.223	**0.685**	0.151	0.064
2. This course allocation system is easy for me to use.	**0.262**	**0.323**	-0.056	0.188	-0.179	0.059	-0.225	0.190
3. I understand how this course allocation system works.	0.174	**0.412**	-0.202	**0.545**	0.089	0.250	**0.494**	-0.221
4. This course allocation system led to the best outcome I could hope for.	**0.274**	-0.226	0.261	0.294	-0.160	-0.188	0.091	-0.212
5. I am satisfied with my course outcome.	**0.258**	-0.304	0.318	0.341	-0.030	-0.237	0.015	0.036
6. I enjoyed participating in this course allocation system.	**0.291**	0.053	0.053	-0.147	-0.060	0.303	**-0.447**	-0.120
7. I like this course allocation system.	**0.327**	0.034	-0.017	-0.164	-0.063	0.041	-0.076	-0.304
8. My fellow students will like this course allocation system.	**0.289**	0.142	-0.081	-0.288	-0.107	-0.377	0.271	-0.278
9. I felt like I had control over my schedule in this course allocation system.	**0.299**	-0.046	0.080	-0.124	-0.052	-0.046	-0.171	**-0.381**
10. This course allocation system is simple.	0.227	**0.400**	-0.290	0.121	-0.063	-0.188	-0.350	0.294
11. I had to think strategically about what other students would do in this course allocation system.	-0.090	**0.364**	**0.638**	-0.127	**-0.545**	0.166	0.168	0.191
12. Someone with perfect knowledge of the historical supply and demand for courses could have had an advantage over me in this system.	-0.120	**0.399**	**0.502**	0.048	**0.676**	-0.132	-0.203	-0.195
Final Survey Questions								
13. Which course allocation system did you prefer?	**0.323**	-0.016	0.022	-0.154	0.143	0.070	-0.002	0.154
14. Which course allocation system do you think your fellow students would prefer?	**0.268**	0.135	0.005	**-0.439**	0.233	-0.200	**0.418**	0.335
15. In which course allocation system did you get a better schedule?	**0.280**	-0.222	0.144	0.238	0.173	-0.104	-0.029	**0.488**
Summary Principal Component Information								
Component Standard Deviation	2.802	1.219	1.073	0.907	0.819	0.724	0.714	0.655
Proportion of Total Variance	0.523	0.099	0.077	0.055	0.045	0.035	0.034	0.029
Cumulative Proportion of Total Variance	0.523	0.622	0.699	0.754	0.799	0.834	0.868	0.896

Notes: Table A6 provides the correlation coefficients for the first eight principal components. The variables included in the analysis are the 15 survey questions, treating the first 12 survey questions as difference variables (the A-CEEI response less the Auction response) and leaving the last three survey questions as is. Bolded values are greater than or equal to .75 times the maximum coefficient magnitude for that column.